UNIVERSIDAD AUTÓNOMA
METROPOLITANA-IZTAPALAPA

DIVISIÓN DE CIENCIAS BÁSICAS E INGENIERÍAS

Definición y Evaluación del Desempeño de un Biomarcador para el Tratamiento de Cáncer de Mama

TESIS

PARA OBTENER EL GRADO DE:

Maestro en Ciencias
(Matemáticas Aplicadas e Industriales)

PRESENTA:

Alan Jiménez Balandra

ASESOR:

Dr. Gabriel Escarela Pérez

CIUDAD DE MÉXICO, MAYO DE 2016

Índice general

Introducción

El cáncer de mama es el segundo cáncer más común a nivel mundial y el más frecuente en las mujeres con un estimado de 1.67 millones de nuevos casos diagnosticados en 2012 (25 % de todos los casos diagnosticados de cáncer pertenecen a este tipo especifico) [1]. El seguimiento posoperatorio y la eventual cura representan importantes problemas de salud epidemiológica, social y pública, lo cual genera una carga económica considerable para los estados. En este estudio se pretende estimar las probabilidades de supervivencia de mujeres diagnosticadas con cáncer de mama y que se sometieron a algún tratamiento, ya sea quirúrgico o de radioterapia y considerando dos tipos de muerte: muerte por cáncer de mama o por otras causas. Es importante la investigación sobre la mortalidad a largo plazo distinguiendo ambas causas, esto permite la separación de la población en los pacientes que finalmente mueren de cáncer de mama y de aquellos que mueren de otras causas, ayudando a mejorar la precisión de la estimación de la tasa de riesgo para cada subpoblación y separando los efectos que tienen las variables predictoras (raza, edad, tamaño del tumor entre otros) en cada causa de muerte. Un enfoque que considere estas variables predictoras puede ayudar a identificar a los pacientes con un riesgo elevado de muerte eventual de cáncer de mama aumentando la eficacia de las estrategias de tratamiento dirigidas y mejorando la relación costo beneficio, a diferencia de estudios que se centran en un seguimiento a corto plazo, en la supervivencia sin tomar en cuenta estas distinciones o bien, en la supervivencia solamente tomando en cuenta las muertes por cáncer de mama.

Se utilizó una base de datos con 16511 registros de mujeres diagnosticadas con cáncer de mama en 1990 y que se sometieron a alguna cirugía, el final del seguimiento se terminó el 31 de Diciembre del 2011. Esta base se generó a partir de datos del programa Surveillance, Epidemiology, and End Results (SEER) del National Cancer Institute de Estados Unidos, el cual es un sistema de vigilancia epidemiológica de alta calidad que genera datos de incidencia y de supervivencia para pacientes de cáncer diagnosticado. Los registros de este programa recolectan rutinariamente información sobre pacientes residentes de Alaska, California, Connecticut, Georgia, Hawaii, Iowa, Michigan, Nuevo México, Utah y Washington -en estos estados se concentra el 26 % del total de la población de Estados Unidos de América-.

Se usó un modelo de mezclas de decremento múltiple para identificar factores y va-

riables confusas que afectaron ya sea en la supervivencia eventual por causas específicas o la condicional para cada causa de muerte o ambos. Interpretaciones útiles de los coeficientes correspondientes a las relaciones curvilíneas de edad al diagnóstico y factores sociodemográficos y clínico-patológicas ayudaron a profundizar en la comprensión de los riesgos por causas específicas subyacentes y las proporciones de riesgo condicionales. La metodología empleada en el estudio permitió la exploración y definición de un biomarcador predictivo en términos de características importantes de un paciente, el cual mostró ser una herramienta precisa tanto para la cuantificación de la severidad de la enfermedad como para la discriminación de pacientes con diferentes niveles de riesgo de eventualmente morir de cáncer de mama.

El lector encontrará, en el primer capítulo, temas médicos relacionados al cáncer de mama que nos servirán para entender tanto la base de datos, como el biomarcador predictivo. En el capítulo dos se hace un breve resumen de conceptos estadísticos y matemáticos indispensables para comprender el modelo que se utilizó así como la forma en que fue ajustado a los datos. En el capítulo tres nos enfocamos al tratamiento de datos faltantes, en especial el método MICE que será usado en la aplicación de este trabajo. Por último en el capítulo cuatro se verá la aplicación a la base de datos haciendo uso del modelo que se presentó en el capítulo dos y tratando los datos faltantes como en el capítulo tres.

Objetivos

Los objetivos de este trabajo son:

- Proponer un modelo de mezclas para el análisis de riesgos competitivos que permita incluir variables explicativas tanto demográficas como aquellas que conciernen al estado de salud del paciente.
- Emplear una metodología que permita incluir datos de pacientes cuyos registros presentan al menos una variable no observada.
- Desarrollar una metodología basada en verosimilitud para incluir variables explicativas incompletas.
- Obtener el marcador predictivo en forma de una combinación lineal de variables explicativas con los coeficientes de la regresión logística del componente de ponderaciones del modelo de mezclas de las causas de muerte.
- Emplear herramientas de diagnósticos para criticar la bondad de ajuste del modelo propuesto.

Capítulo 1

Cáncer de mama

En este capítulo se abordaran temas relacionados con el cáncer de mama, que es el cáncer de mama, donde se presentan, las causas que lo generan, los tratamientos de esta enfermedad y demás términos médicos importantes que son necesarios para el objetivo de este trabajo.

1.1. Conceptos Médicos

Marcador predictivo: Se define como un marcador que puede ser usado para identificar subpoblaciones de pacientes que son más probables a responder a un tratamiento dado. Con los marcadores predictivos es posible seleccionar el tratamiento con la más alta probabilidad de eficacia a un paciente, de este modo, los marcadores predictivos son la base para el tratamiento individualizado.

Perimenopausia: También llamada premenopausia, es el periodo de transición natural hacia la menopausia. Es la etapa de la vida de una mujer en la que empieza a disminuir la reserva ovárica y aparecen irregularidades en el ciclo menstrual. Suele presentarse entre los 40 y los 48 años de edad. La duración promedio de la perimenopausia suele ser de 4 años, pero para algunas mujeres puede ser de unos meses o de 10 años, básicamente abarca hasta un año después de la menopausia, que es cuando una mujer ha estado 12 meses sin tener el periodo.

Menopausía: Se define como el cese permanente de la regla o menstruación. Suele comenzar en torno a los cuarenta y cinco años, aunque varía bastante de una mujer a otra y no se produce de forma repentina, si no que se trata de un cambio progresivo ya que con el tiempo los ovarios pierden gradualmente la habilidad de producir estrógeno y progesterona, las hormonas que regulan el ciclo menstrual y debido a esto la ovulación y menstruación ocurren con menos frecuencia y eventualmente se detienen.

Cáncer: El cuerpo humano está compuesto por millones de millones de células vivas. Las células normales del cuerpo crecen, se dividen para crear nuevas células y mueren de manera ordenada. Durante los primeros años de vida de una persona, las células normales se dividen más rápidamente para permitir el crecimiento. Una vez que

la persona alcanza la edad adulta, la mayoría de las células se dividen sólo para reponer aquellas que se han desgastado, dañado o muerto. El cáncer se origina cuando las células en alguna parte del cuerpo comienzan a crecer de manera descontrolada.

El crecimiento de las células cancerosas es diferente al crecimiento de las células normales. En lugar de morir, las células cancerosas continúan creciendo y formando nuevas células cancerosas (Figura 1.1). En la mayoría de los casos, las células cancerosas forman un tumor. Las células cancerosas también pueden crecer hacia otros tejidos (invadir), algo que las células normales no hacen. La posibilidad de una célula de crecer sin control e invadir otro tejido es lo que la hace cancerosa.

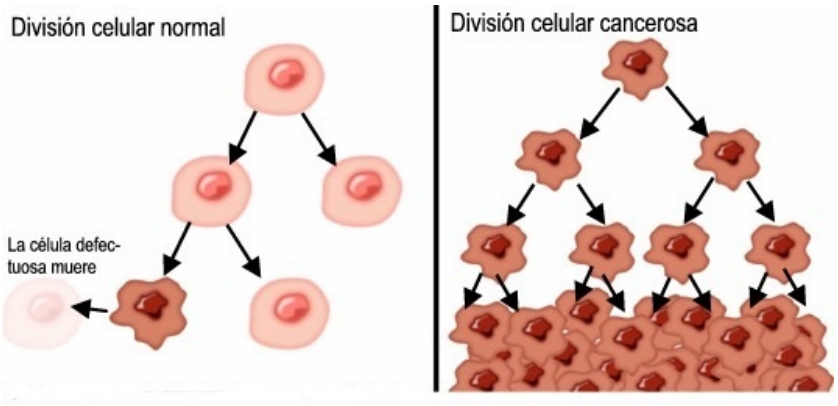


Figura 1.1: Células normales y células cancerígenas

Los diferentes tipos de cáncer se pueden comportar de manera muy distinta. Cada tipo de cáncer crece a velocidad distinta y responde a distintos tratamientos. Es por esto que las personas con cáncer necesitan recibir un tratamiento dirigido a su propio tipo de cáncer.

No todos los tumores son cancerosos. A los tumores que no son cancerosos se les llama tumores benignos. Los tumores benignos pocas veces ponen en riesgo la vida de una persona pero pueden causar problemas, ya que pueden crecer mucho y ocasionar presión en los tejidos y órganos sanos. Sin embargo, estos tumores no pueden crecer hacia otros tejidos, por esta razón, no pueden propagarse hacia otras partes del cuerpo (no pueden hacer metástasis).

Metástasis: Algunas veces las células cancerosas se propagan a otras partes del cuerpo. En este nuevo sitio, pueden continuar creciendo y formar nuevos tumores. A este proceso se le conoce como metástasis.

Independientemente del lugar hacia el cual se propague el cáncer, se le da el nombre (y se trata) según el lugar donde se originó. Por ejemplo, el cáncer de próstata que se extendió a los huesos sigue siendo cáncer de próstata y no cáncer de huesos.

Cáncer de mama: El cáncer de mama es un tumor maligno que se inicia en las células de la mama. Un tumor maligno es un grupo de células cancerosas que pueden desarrollarse en los tejidos o propagarse a áreas distantes del cuerpo de los alrededores.

La enfermedad se presenta casi exclusivamente en las mujeres, pero los hombres también pueden tenerlo.

El seno de una mujer está formado por glándulas que pueden producir leche (lobulillos), pequeños conductos que llevan la leche de los lobulillos al pezón (ductos), tejido adiposo y conectivo, vasos sanguíneos y vasos linfáticos. La mayoría de los cánceres de seno comienza en las células que revisten a los conductos. Otros cánceres menos frecuentes de seno se originan en las células que revisten los lobulillos. Los cánceres también pueden comenzar en las células de otros tejidos en el seno, a estos se les llama sarcomas y linfomas, y en realidad no se consideran cáncer de mama. La siguiente figura muestra como esta estructurado un seno:

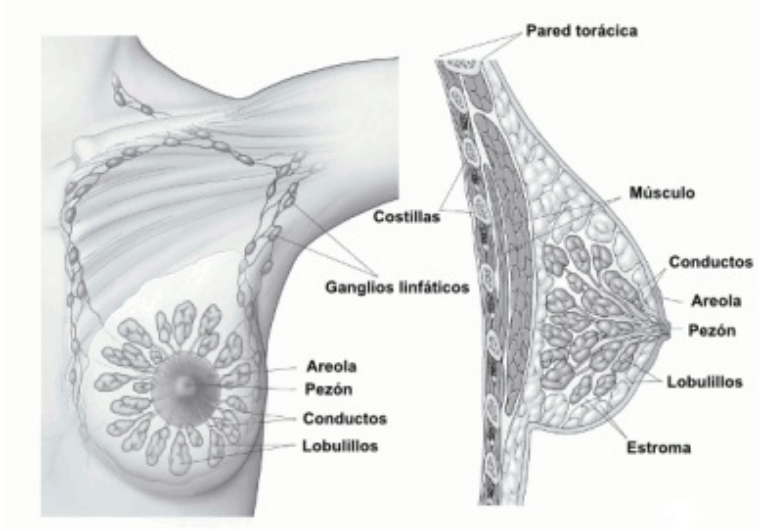


Figura 1.2: Partes del seno

Una de las principales maneras de propagación del cáncer de seno es a través del sistema linfático. Normalmente, los ganglios linfáticos son pequeños grupos en forma de fríjol de tejidos que contiene cierta clase de célula del sistema inmunológico (células que combaten infecciones). Los ganglios linfáticos están conectados por vasos (como pequeñas venas) que transportan un líquido claro llamado linfa en lugar de sangre.

La mayoría de los vasos linfáticos del seno drenan hacia los ganglios linfáticos localizados debajo del brazo (ganglios axilares), los ganglios linfáticos alrededor de la clavícula (ganglios linfáticos infraclaviculares y ganglios linfáticos supraclaviculares) o hacia los ganglios linfáticos que se encuentran en el interior del tórax y cerca del esternón (ganglios linfáticos mamarios internos). La siguiente figura muestra como estan localizados los ganglios linfaticos:

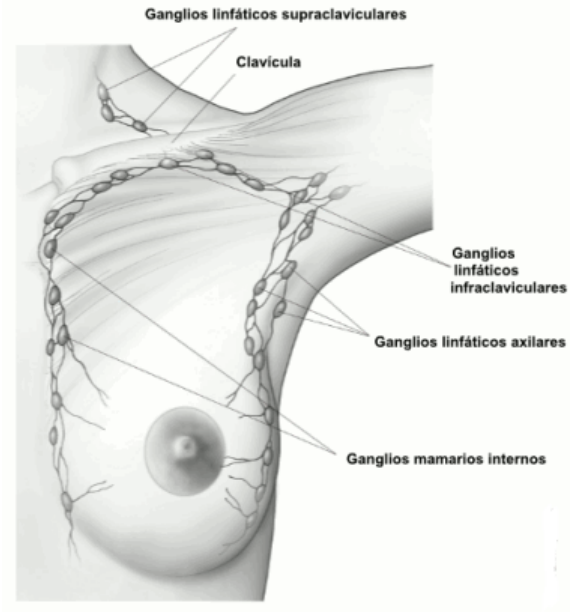


Figura 1.3: Ganglios linfaticos en el seno

Receptor de hormonas: Son proteínas especiales que se encuentran en el interior y sobre la superficie de ciertas células mamarias. Estas proteínas receptoras funcionan como especies de ojos y oídos para las células, ya que reciben mensajes de sustancias que circulan por el torrente sanguíneo y les indican qué hacer. En otras palabras, los receptores actúan como un interruptor que activa o desactiva una función particular en la célula. Si la sustancia adecuada se acopla al receptor, como una llave que encaja en una cerradura, el interruptor se activa e inicia una función específica de la célula (Figura 1.4).

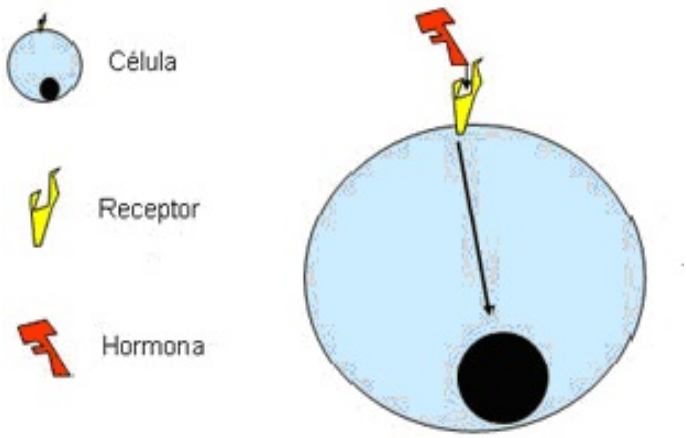


Figura 1.4: Representación de los receptores de hormona

Al unirse a receptores de hormonas, el estrógeno y la progesterona contribuyen

al desarrollo y funcionamiento de las células mamarias. El estrógeno y la progesterona cumplen una función importante en el ciclo menstrual, el desarrollo sexual, el embarazo y el parto. Pero aun después de la menopausia, las mujeres siguen teniendo estas hormonas en el cuerpo. Los hombres también las tienen, aunque en cantidades mucho menores.

Tal como las células mamarias sanas, la mayoría de las células cancerígenas —pero no todas— tienen receptores de hormonas y responden a las señales emitidas por esas hormonas. Es importante saber si las células cancerígenas de las mamas tienen receptores de hormonas o no para tomar decisiones de tratamiento. Si hay células cancerígenas en las mamas con receptores de hormonas, se puede emplear hormonoterapia para interrumpir la influencia de las hormonas en el crecimiento y el funcionamiento general de las células. Si se elimina o se bloquea la hormona, tal como lo hacen estos medicamentos, las células cancerígenas tienen menos probabilidades de sobrevivir.

También es importante destacar que algunos tipos de cáncer de mama de receptores de hormonas positivos pueden perder esos receptores con el paso del tiempo. También sucede lo opuesto: los tipos de cáncer sin receptores de hormonas pueden pasar a llegar a tenerlos. Si las células cancerígenas ya no tienen receptores, es poco probable que la hormonoterapia sirva para tratar el cáncer. En cambio, si las células han adquirido receptores de hormonas, la hormonoterapia sí puede ser útil.

Los laboratorios hacen un estudio de una muestra de tejido mamario canceroso para ver los receptores de hormona, no todos los laboratorios usan el mismo método para analizar los resultados de la prueba ni informan sobre los resultados de la misma manera. Una manera de informar al paciente los resultados es por medio de las categorías “positivo”, “negativo” o “al limite”, y diferentes laboratorios establecen distintos valores para considerarlos en estas categorías; por ejemplo, si menos del 10 % de las células dan positivo, este podría ser un resultado negativo para un laboratorio mientras que otro lo podría considerar positivo.

Tipo de cáncer de mama:

Los tipos de cáncer de mama son clasificados dependiendo de donde se desarrolló el cáncer y si se propago o no. Los que no se propagan hacia otros tejidos mamarios son llamados -in situ- y los que se han propagado -invasivo o infiltrante-. Con respecto al lugar en que se desarrollan existen dos tipos principales:

- El carcinoma ductal: comienza en los conductos que llevan leche desde los lobulillos hasta el pezón. Este tipo de cáncer de mama es el más común, ya que se estima que uno de cada cinco casos nuevos de cáncer de mama serán carcinoma ductal in situ (también conocido carcinoma intraductal) y el carcinoma ductal invasivo es aproximadamente el 80 % de todos los cánceres de mama invasivos.

- El carcinoma lobulillar: este tipo de cáncer comienza a crecer en los lobulillos que son las glándulas productoras de leche ubicadas en los extremos de los conductos mamarios. El carcinoma lobulillar invasivo es el segundo cáncer invasivo más común, con un estimado del 10 % con respecto a todos los cánceres invasivos. El carcinoma lobulillar in situ - también llamado neoplasia lobulillar in situ- no es un verdadero cáncer de mama sino una señal de que la persona presenta un riesgo

mayor al promedio de padecerlo en el futuro. La neoplasia es una acumulación de células anómalas.

Hay otros tipos o subtipos especiales de los anteriores que son menos comunes -con un porcentaje menor al 5 % de todos los cánceres-, algunos de ellos son:

- Carcinoma mucinoso: Se le conoce también como carcinoma coloide. Es un tipo de cáncer invasivo de seno formado por células cancerosas que producen mucosidad.
- Carcinoma medular: Es un cáncer invasivo de seno que se caracteriza por tener un límite bien establecido entre el tejido del tumor y el tejido normal. Las células cancerosas se presentan en gran tamaño y en los bordes del tumor existen células del sistema inmune.
- Cáncer inflamatorio de seno: Este tipo de cáncer suele no presentarse como un sólo tumor. Hace que la piel del seno tenga un aspecto rojizo y grueso con hoyuelos similares a la cáscara de una naranja, volviéndose más grande, firme, sensible y puede sentirse picazón y calor en la zona debido a que las células cancerosas bloquean los vasos linfáticos de la piel. Suele tener una mayor probabilidad de propagación y un peor pronóstico que el cáncer ductal invasivo o lobulillar invasivo.
- Carcinoma papilar: Puede ser invasivo o no invasivo. Se considera un subtipo de carcinoma ductal in situ. Las células de estos cánceres tienden a estar agrupadas en proyecciones pequeñas que se asemejan a los dedos. Tienden a ser diagnosticados a mujeres de edad avanzada.
- Carcinoma metaplásico: También conocido como carcinoma con metaplasia, es un tipo de cáncer ductal invasivo que se caracterizan por tener células que normalmente no se encuentran en el seno, como células de la piel o células de los huesos.
- Tumores mixtos: Contienen una variedad de tipos de células, tal como células del cáncer ductal invasivo combinadas con células del cáncer lobulillar invasivo. El tratamiento en contra de este tipo de cáncer es de igual forma que el cáncer ductal invasivo.
- Tumor filoides: Otros nombre que se le suele denominar es cistosarcoma filoides, este tipo de cáncer se forma en el estroma del seno. Por lo general, estos tumores son benignos pero en pocos casos son malignos. Para su tratamiento es necesario extirparlo junto con un borde más amplio de tejido normal.
- Enfermedad de Paget del pezón: Comienza en los conductos del seno y se propaga hacia la piel del pezón y a la areola.Como consecuencia aparecen costras, escamas, áreas de sangrado o supuración y luce enrojecida, en ocasiones se experimenta ardor o comezón.

- Carcinoma quístico adenoide: Llamado también carcinoma adenoquístico, tiene características glandulares (adenoides) y con apariencia cilíndrica, en algunas ocasiones se propagan a los ganglios linfáticos o áreas distantes.

- Carcinoma tubular: Es un tipo especial del carcinoma ductal invasivo, se les da el nombre de tubulares por la manera en que las células están agrupadas, es tratado como un carcinoma ductal invasivo.

Grado del tumor:
El grado del tumor de seno refleja qué tan normales se observan sus células y en general el tejido que lo conforma. También indica qué tan rápido crece y se disemina, es decir, indica qué tan agresivo es. Si existe la sospecha de que un tumor es maligno es necesario realizar una biopsia, es decir remover todo o una parte de este para que un patólogo examine su tejido y así determine si es o no maligno y el grado en el que se encuentra.

Existen distintos sistemas que determinan el grado del tumor de seno. Uno de ellos es el Sistema Nottingham Histologic Score, en el cual se toman en cuenta tres factores: la ”diferenciación”que se refiere a qué tan bien las células tumorales recrean las 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 un score de 1 a 3 que luego se suma para obtener un score final de 3 a 9, y que determina el grado del tumor de la siguiente forma:

- El score final de 3 a 5 corresponde al grado 1 también llamado bien diferenciado.

- El score final de 6 a 7 corresponde al grado 2 o al moderadamente diferenciado.

- El score final de 8 a 9 corresponde a los grados 3 y 4 llamados pobremente diferenciado y no diferenciado.

Si las células del tumor de seno y su organización en el tejido se observan casi normales entonces se dice que el tumor está bien diferenciado, mientras que un tumor con células y tejido claramente anormales se llama pobremente diferenciado. Los tumores de grados 3 y 4 crecen y se esparcen más rápido que los de grados más bajos.

1.2. Tratamiento

En esta sección se presentan distintos tipos de tratamiento que son utilizados para contrarrestar o atacar el cáncer de mama.

Cirugía
La cirugía es utilizada para el diagnóstico o como tratamiento para curar una enfermedad. Para el caso específico del cáncer de mama es utilizado para examinar el tipo

de cáncer, si hay afectación en los nodos linfáticos o extirpar el tumor. Se mostraran los tipos de cirugía utilizados para el tratamiento del cáncer de mama.

Cirugía con conservación del seno: A veces se le llama mastectomía parcial, segmentaria, tumorectomía y cuadrantectomía. En este tipo de cirugía sólo se elimina la parte de la mama que contiene el cáncer así como parte del tejido normal circundante, una ilustración de esto se muestra en la figura 1.5. La porción de la mama que es eliminada depende de factores del tumor como el tamaño, la ubicación, entre otros. Es probable el cambio de la forma de la mama después de la cirugía, por lo tanto, es posible tener algún tipo de cirugía reconstructiva, o para que el tamaño de la mama no afectada sea reducido para hacer que los senos sean más simétricos.

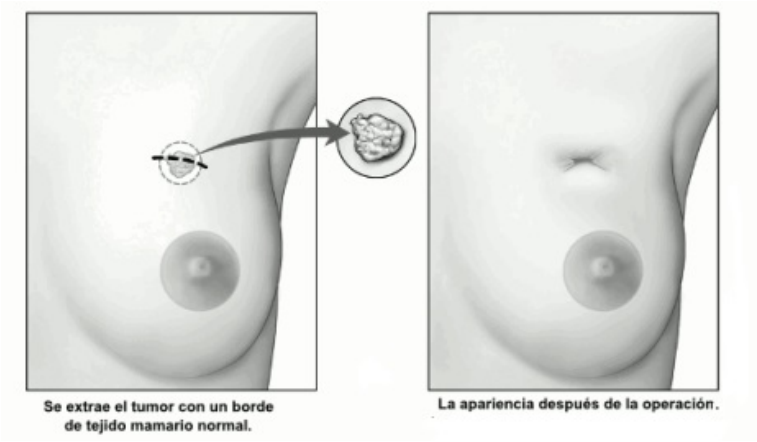


Figura 1.5: Cirugía con conservación del seno

Mastectomía

La mastectomía es una cirugía para extirpar todo el tejido mamario a veces junto con otros tejidos cercanos (figura 1.6). Entre los tipos más comunes de mastectomía esta:

- Mastectomía simple: Es el tipo más común de la mastectomía que se utiliza para tratar el cáncer de mama. En este procedimiento se extirpa todo el seno, incluyendo el pezón, pero no elimina los ganglios linfáticos de la axila o el tejido muscular de debajo de la mama. A veces se extirpan ambos senos como cirugía preventiva en mujeres con alto riesgo de contraer cáncer en el otro seno. La mayoría de las mujeres, si son hospitalizadas, pueden ir a casa al día siguiente.

- Mastectomía conservadora de piel: En este procedimiento la mayor parte de la piel sobre el pecho (que no sea el pezón y la areola) se deja intacta. La cantidad de tejido mamario eliminado es el mismo que con una mastectomía simple. Se usa cuando se planea la reconstrucción mamaria inmediata utilizando implantes o tejidos de otras partes del cuerpo. No es adecuado para tumores grandes o los que están cerca de la superficie de la piel.

- Mastectomía con conservación del pezón: En este procedimiento se elimina el tejido del seno pero la piel de la mama y el pezón se deja en su lugar, al igual que la mastectomía conservadora de piel es utilizado para la reconstrucción mamaria.
- Mastectomía radical: En esta operación se extirpa todo le seno, los ganglios linfáticos axilares y el pectoral así como los músculos debajo del pecho. Sólo es realizada cuando existen grandes tumores que crecen en los músculos pectorales debajo del seno.
- Mastectomía radical modificada: Es una mastectomía simple con eliminación de algunos ganglios linfáticos de la axila.

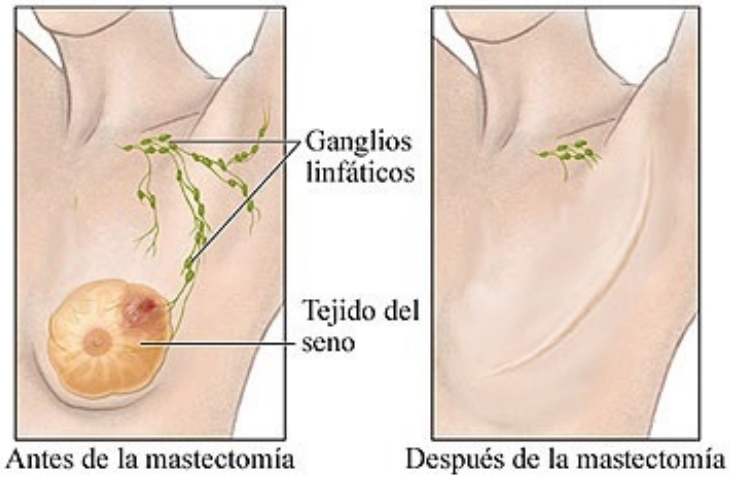


Figura 1.6: Mastectomía

Radioterapia

La radioterapia es un tratamiento con rayos de alta energía o partículas que destruyen las células cancerosas. La radiación en la mama a menudo se administra después de la cirugía de preservación de mama para ayudar a reducir la probabilidad de que el cáncer regrese al seno o a los ganglios linfáticos cercanos, también puede ser recomendada después de la mastectomía en pacientes con un cáncer de más de 5 cm, o cuando el cáncer se encuentra en los ganglios linfáticos. Es utilizado para tratar el cáncer metastásico. Es un procedimiento indoloro y suele durar unos minutos. La medida de radiación depende del tipo de cirugía que se llevo a cabo o si los ganglios linfáticos están afectados. Cuando se administra después de la cirugía, la radioterapia por lo general no se inicia hasta que los tejidos han sido capaces de curarse.

Capítulo 2

Métodos estadísticos

En este capitulo se dan algunos métodos y teoría estadística que serán necesarios en la aplicación a la base de datos de cáncer de mama. Se iniciará con la definición de polinomios ortogonales los cuales se ajustaran a la variable edad para poder definir el biomarcador. Después se presentan algunos métodos estadísticos que fueron utilizados para el ajuste y para el tratamiento de datos faltantes. Por último se presenta una breve introducción al análisis de supervivencia iniciando con el caso de un sólo tipo de fallo (análisis de supervivencia univariado) para así poder generalizarlo a varios tipos de fallo (decremento múltiple) y con estos elementos definir el modelo que se va a utilizar.

2.1. Polinomios ortogonales

Hay dos razones principales por las que la aproximación polinomial es usada en estadística. La primera es modelar relaciones no lineales entre una variable respuesta y una explicativa. Comúnmente la primera es medida con error, y el interés está en la forma de la curva ajustada y quizá también en los coeficientes polinomiales ajustados. Ya que se busca la parsimonia en un modelo así como su interpretación, rara vez se estará interesado en curvas polinomiales de órdenes mayores a tres o cuatro. La segunda razón es aproximar una función que es difícil de evaluar, tal y como son algunas funciones de distribución o densidad. Aquí no interesa la curva polinomial en sí; lo importante es qué tan cerca el polinomio puede seguir dicha función, y especialmente qué tanto se puede reducir el error. Pueden usarse polinomios de orden alto si ellos proveen aproximaciones exactas. Generalmente el primer paso al aproximar una función es transformarla o estandarizarla para hacerla manejable para la aproximación polinomial. En cualquiera de los dos casos los polinomios ortogonales ofrecen características deseables como coeficientes no correlados lo que a su vez permite minimizar el error de aproximación y la sensibilidad de los cálculos de redondeo.

Supóngase que se quiere aproximar la función $f(x)$ y que ésta se observa en los valores $x_1, x_2, \ldots, x_N$. Se quiere estimar $f(x)$ para nuevos valores de x. Si esos valores están dentro del rango de las x's observadas entonces el problema se llama interpolación,

pero si están fuera se llama extrapolación. Se ha visto que los polinomios son útiles en la interpolación y muy deficientes en la extrapolación.

El polinomio general $p(x) = c_0 + c_1x + \ldots + c_nx^n$ está escrito en términos de los monomios x^j. El problema es que los monomios se ven casi igual cuando se grafican en el intervalo [0,1], es decir están muy correlados. Esto significa que pequeños cambios en $p(x)$ podrían ser provocados por cambios relativamente grandes en los coeficientes $c_0, c_1, \ldots, c_n$, los cuales no están bien determinados cuando hay error de redondeo.

El polinomio general también puede escribirse en términos de cualquier secuencia de polinomios básicos de grado creciente $p(x) = a_0p_0(x) + a_1p_1(x) + \ldots + a_np_n(x)$ donde el grado de $p_j(x)$ es j para $j = 0, \ldots, n$. Hay una relación lineal entre los coeficientes originales c_j y los nuevos coeficientes a_j para hacer el polinomio resultante igual en cada caso.

La idea detrás de los polinomios ortogonales es seleccionar los polinomios basicos $p_j(x)$ para que sean diferentes de los otros tanto como sea posible. Dos polinomios p_i y p_j son ortogonales si $p_i(X)$ y $p_j(X)$ están no correlacionados cuando X varía sobre alguna distribución como la Gamma o la Normal estándar. Cualquier sucesión de polinomios ortogonales puede calcularse recursivamente usando una fórmula de tres términos.

2.2. Estimación por el Método de Máxima Verosimilitud

La función de verosimilitud para un modelo estadístico se define de la misma manera que la función de densidad pero se le considera una función de $\boldsymbol{\theta}$ para datos fijos $\mathbf{z}$ de esta forma $L(\boldsymbol{\theta}|\mathbf{z}) = f(\mathbf{z}|\boldsymbol{\theta})$.

El método de máxima verosimilitud se usa para obtener un estimador del verdadero valor de $\boldsymbol{\theta}$ el cual establece que se busque el vector de valores $\hat{\boldsymbol{\theta}}$ mediante la optimización de la función de verosimilitud de manera que se dé la mayor probabilidad de que ocurra el evento de interés dado $\mathbf{z}$. Por lo tanto, el estimador de máxima verosimilitud (EMV) de $\boldsymbol{\theta}$ se define como el valor $\hat{\boldsymbol{\theta}} \in \Theta$ donde Θ es el conjunto de todos los posibles valores del vector de parámetros $\boldsymbol{\theta}$ el cual maximiza $L(\boldsymbol{\theta}|\mathbf{z})$.

Debido a que la función logaritmo natural es monótona, el máximo de la función de $\ell(\boldsymbol{\theta}|\mathbf{z}) = \log[L(\boldsymbol{\theta}|\mathbf{z})]$ llamada log verosimilitud es igual al máximo de la función de verosimilitud, por varias razones es más conveniente usar esta función en lugar de la función de verosimilitud.

Los momentos de la derivada de la función log verosimilitud satisfacen propiedades importantes, entre estas están que la primera derivada tiene esperanza cero y la varianza de la primer derivada es menos la esperanza de la segunda derivada, esto se resume en las siguientes ecuaciones:

$$\mathrm{E}_{\boldsymbol{\theta}}[\nabla \ell(\boldsymbol{\theta}|\mathbf{z})] = 0 \tag{2.1}$$

$$\mathrm{Var}_{\boldsymbol{\theta}}[\nabla \ell(\boldsymbol{\theta}|\mathbf{z})] = -\,\mathrm{E}_{\boldsymbol{\theta}}[\nabla^2 \ell(\boldsymbol{\theta}|\mathbf{z})] \tag{2.2}$$

donde $\nabla \ell(\boldsymbol{\theta}|\mathbf{z})$ y $\nabla^2 \ell(\boldsymbol{\theta}|\mathbf{z})$ denotan el gradiente y el hessiano de $\ell(\cdot)$ evaluado en $\mathbf{z}$, respectivamente. A la ecuación (2.2) se le llama información de Fisher esperada y se denota por $\mathrm{I}(\boldsymbol{\theta})$.

La aproximación asintótica del EMV es una normal multivariada con media $\boldsymbol{\theta}$ (el valor verdadero del parámetro) y varianza $\mathrm{I}(\boldsymbol{\theta})^{-1}$. Es posible aproximar la varianza asintótica usando el EMV, es decir, se utiliza $\mathrm{I}(\hat{\boldsymbol{\theta}}_{\mathbf{z}})^{-1}$ como la varianza aproximada del EMV. La teoría asintótica garantiza que esta aproximación produce un error que es despreciable comparado con la mejor aproximación de la distribución del EMV dado por la $N(\boldsymbol{\theta}, \mathrm{I}(\boldsymbol{\theta})^{-1})$.

2.3. Prueba estadística de Wald

Una prueba estadística es un procedimiento que nos permite evaluar hasta que punto una hipótesis es consistente. En la prueba estadística de Wald, la estimación de máxima verosimilitud $\hat{\theta}$ del parámetro de interés θ se compara con el valor propuesto θ_0, bajo la suposición de que la diferencia entre ambos seguirá aproximadamente una distribución normal. En el caso univariado, el estadístico de Wald es:

$$\frac{(\hat{\theta} - \theta_0)^2}{\mathrm{var}(\hat{\theta})} \tag{2.3}$$

el cual se compara con una distribución χ^2_1. Alternativamente, la diferencia puede ser comparada con una distribución normal estándar. En este caso el resultado el estadístico de Wald es:

$$\frac{\hat{\theta} - \theta_0}{\mathrm{e.\,e.}(\hat{\theta})}$$

2.4. Método de Monte Carlo vía Cadenas de Markov

Sea $\boldsymbol{Y}$ un conjunto de datos observados, y una distribución a priori para un conjunto de parámetros $\boldsymbol{\phi}$ dada por $f(\boldsymbol{\phi}|\boldsymbol{\theta})$, donde $\boldsymbol{\theta}$ es una distribución a priori elegida por el investigador. Entonces, por el teorema de Bayes, la distribución a posteriori de los parámetros dado los datos es dado por:

$$f(\boldsymbol{\phi}|\boldsymbol{Y}) = \frac{f(\boldsymbol{Y}|\boldsymbol{\phi}) f(\boldsymbol{\phi})}{\int f(\boldsymbol{Y}|\boldsymbol{\phi}) f(\boldsymbol{\phi})\, \mathrm{d}\boldsymbol{\phi}} \propto f(\boldsymbol{Y}|\boldsymbol{\phi}) f(\boldsymbol{\phi})$$

Para ciertos casos la distribución a priori puede calcularse sin ningún problema, pero comúnmente no es factible obtenerla analíticamente. Para esto existe una técnica llamada

método de Monte Carlo vía Cadenas de Markov (denotada por MCMC por su nombre en inglés Markov Chain Monte Carlo) el cual es un muestreo iterativo en que se obtiene una secuencia (conocida como cadena) de valores de los parámetros. Esta cadena que se representará por $\boldsymbol{\phi}^{(0)}, \boldsymbol{\phi}^{(1)}, \ldots$ tiene la propiedad que su distribución estacionaria es la distribución a posteriori $f(\boldsymbol{\phi}|\boldsymbol{Y})$. Así corriendo este algoritmo con un valores inicial y desechando los primeros valores obtenidos en la cadena, los valores restantes son una muestra correlada de la distribución a posteriori, de la que es posible estimar su media, moda o varianza, mediante dicha muestra. Existen dos algoritmos mediante los cuales se puede obtener la cadena $\boldsymbol{\phi}^{(0)}, \boldsymbol{\phi}^{(1)}, \ldots$, el algoritmo Metropolis – Hastings y el muestreo de Gibbs, donde el segundo es un caso especial del primero.

2.4.1. Muestreo Metropolis-Hasting

Supóngase que $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_p)^T$ es el vector de parámetros de interés y defínase $\boldsymbol{\phi}_{-j} = (\phi_1, \phi_2, \ldots, \phi_{j-1}, \phi_{j+1}, \ldots, \phi_p)^T$ el vector $\boldsymbol{\phi}$ sin el parámetro ϕ_j. El algoritmo 1 muestra este método.

Algoritmo 1 Muestreo Metropolis-Hasting

1: Definir los valores iniciales para cada entrada del vector $\boldsymbol{\phi}$, que se denotará por $\boldsymbol{\phi}^{(0)}$.
2: Actualizar el paso $r = 1, 2, \ldots$ haciendo $\boldsymbol{\phi}^{(r)} = \boldsymbol{\phi}^{(r-1)}$.
3: **for** $i = 1 \ldots, p$ **do**
4: Muestrear un nuevo valor para $\phi_j^{(r)}$ de una distribución $\tilde{\phi}_j^{(r)} \sim f(\phi|\phi_j^{(r)}, \boldsymbol{\phi}_{-j}^{(r)})$.
5: Calcular la probabilidad de aceptación:

$$p = \text{mín}\left(1, \frac{\{f(\boldsymbol{Y}|\tilde{\phi}_j^{(r)}, \boldsymbol{\phi}_{-j}^{(r)}) f(\tilde{\phi}_j^{(r)}, \boldsymbol{\phi}_{-j}^{(r)})\} f(\phi_j^{(r)}|\tilde{\phi}_j^{(r)}\boldsymbol{\phi}_{-j}^{(r)})}{\{f(\boldsymbol{Y}|\phi_j^{(r)}, \boldsymbol{\phi}_{-j}^{(r)}) f(\phi_j^{(r)}, \boldsymbol{\phi}_{-j}^{(r)})\} f(\phi_j^{(r)}|\phi_j^{(r)}\boldsymbol{\phi}_{-j}^{(r)})}\right)$$

entonces se saca $a \sim$ uniforme$[0, 1]$ y se acepta $\tilde{\phi}_j^{(r)}$ si $a < p$.
6: Si el valor propuesto es aceptado, se remplaza el valor actual de $\phi_j^{(r)}$ en $\boldsymbol{\phi}^{(r)}$ con el valor propuesto $\tilde{\phi}_j^{(r)}$, si no se sigue dejando el valor actual.
7: **end for**

Si en el paso 4 del algoritmo 1 es posible proponer que la distribución para ϕ_j sea $f(\phi_j|\boldsymbol{\phi}_{-j}, \boldsymbol{Y})$, entonces la probabilidad de aceptación es siempre igual a 1. Este caso especial de distribución propuesta es conocida como el muestreo de Gibbs.

2.5. Regresión Logística Multinomial

Se considera una variable aleatoria Y con J categorías y $p_1, p_2, \ldots, p_J$ sus respectivas probabilidades, con $p_1 + p_2 + \cdots + p_J = 1$. Sea n el número de observaciones independientes de Y con y_1 resultados en la categoría 1, y_2 resultados en la categoría 2, ...,

y_J resultados en la categoría J, entonces se define $\mathbf{y} = [y_1, y_2, \ldots, y_J]^T$. La distribución multinomial es:

$$f(\mathbf{y}|n) = \frac{n!}{y_1!y_2! \cdot y_J!} p_1^{y_1} p_2^{y_2} \cdots p_J^{y_J}$$

Para datos nominales una categoría es elegida arbitrariamente como la categoría de referencia. Sea $\mathbf{z}$ un vector de covariables, si se elige la primer categoría como categoría base entonces se tiene:

$$\text{logit}(p_j) = \log\left(\frac{p_j}{p_1}\right) = \delta_j + \boldsymbol{\delta}_j \mathbf{z} \qquad j = 2, \ldots, J. \tag{2.4}$$

Estas son llamadas las ecuaciones logit. Las $(J-1)$ ecuaciones logit son usadas simultáneamente para estimar los parámetros δ_j y $\boldsymbol{\delta}_j$. De la ecuación (2.4) se tiene $\hat{p}_j = \hat{p}_1 \exp(\delta_j + \boldsymbol{\delta}_j \mathbf{z})$ para $j = 2, \ldots, J$, ya que $\hat{p}_1 + \hat{p}_2 + \cdots + \hat{p}_J = 1$ se deduce:

$$p_1 = \frac{1}{1 + \sum_{l=2}^{J} \exp(\delta_l + \boldsymbol{\delta}_l \mathbf{z})}$$

y para $j = 2, \ldots, J$

$$p_j = \frac{\exp(\delta_j + \boldsymbol{\delta}_j \mathbf{z})}{1 + \sum_{l=2}^{J} \exp(\delta_l + \boldsymbol{\delta}_l \mathbf{z})} \tag{2.5}$$

Este modelo es llamado el **modelo de regresión logística nominal**.

2.6. Criterio de Información Bayesiano (BIC)

El criterio de información bayesiano (BIC) propuesto por Schwarz en (1978) es un criterio para la selección de modelos dentro de un conjunto finito de modelos, en esta sección se presenta un breve introducción para la aplicación de este criterio.

Supóngase que se tienen s modelos no necesariamente anidados y sea $\boldsymbol{Y}$ un conjunto de datos de tamaño n. La densidad condicional dado por el i-ésimo modelo candidato (M_i) y su correspondiente vector de parámetros $(\boldsymbol{\theta}^i)$ esta dada por $f(\boldsymbol{Y}|M_i, \boldsymbol{\theta}^i)$ con $\boldsymbol{\theta}^i \in \Theta_i \subset \mathbb{R}^k$ es el conjunto de los distintos valores que puede tomar el vector de parámetros.

Sea $\pi(\boldsymbol{\theta}^i)$ la densidad a priori para el vector de parámetros $\boldsymbol{\theta}^i$ dado el modelo M_i, $p(M_i)$ una densidad de probabilidad discreta a priori que asigna probabilidad positiva a cada uno de los modelos $M_1, \ldots, M_s$. Por el teorema de Bayes de probabilidad total, la probabilidad a posteriori del i-ésimo modelo es:

$$p(M_i|\boldsymbol{Y}) = \frac{p(M_i) f(\boldsymbol{Y}|M_i)}{f(\boldsymbol{Y})} = \frac{p(M_i) \displaystyle\int_{\Theta_i} f(\boldsymbol{Y}|M_i, \theta^i) \pi(\theta^i) d\theta^i}{f(\boldsymbol{Y})} \tag{2.6}$$

La probabilidad condicional $p(M_i|\boldsymbol{Y})$ se interpreta como la probabilidad de que los datos provengan del modelo M_i dado que se ha observado $\boldsymbol{Y}$, entonces el mejor modelo es el que tenga mayor probabilidad a posteriori.

Ya que el denominador de la ecuación (2.6) es un término común en todos los modelos, solo se utiliza el numerador para comparar modelos, si se supone ademas que las probabilidades a priori $p(M_i)$ son iguales para todos los modelos, entonces el modelo que maximice la integral es el que debe seleccionarse como el mejor. Ya que la integral de la ecuación (2.6) es difícil de calcular, y ademas se necesitan especificar las densidades a priori $\pi(\theta^i)$, se realiza una aproximación del logaritmo de la integral [2]:

$$\log\left(\int_{\Theta_i} f(\boldsymbol{Y}|M_i,\boldsymbol{\theta}^i)\pi(\boldsymbol{\theta}^i)d\boldsymbol{\theta}^i\right) \approx \ell_i(\hat{\boldsymbol{\theta}}^i) - \frac{k_i}{2}\log(n) \tag{2.7}$$

donde $\hat{\boldsymbol{\theta}}^i$, $\ell_i(\cdot)$ y k_i son el estimador de máxima verosimilitud, la log verosimilitud y el número de parámetros, respectivamente, para el modelo M_i. Como la función logaritmo natural es monótona creciente, el que maximice la integral en la ecuación (2.6) equivale aproximadamente a maximizar (2.7), lo que a su vez es el que minimice:

$$BIC = -2\ell_i(\hat{\boldsymbol{\theta}}^i) + k_i\log(n) \tag{2.8}$$

2.7. Variables de confusión

La confusión en un estudio clínico se refiere a la mezcla de los efectos que causa el tratamiento con los efectos causados por uno o varios factores adicionales. Estos últimos, llamados factores de confusión, pueden enmascarar una asociación, o más comúnmente, pueden demostrar equivocadamente una asociación inexistente entre el tratamiento y la consecuencia que se está probando, por lo que para establecer una causalidad se requieren métodos que controlen su efecto. Los factores de confusión compiten con el tratamiento en cuanto a que también proporcionan una explicación al resultado del estudio. En pruebas clínicas es común que la confusión se genere a causa de una distribución desigual de los que podrían ser confusores entre los grupos de tratamiento.

Para que un factor sea de confusión debe existir una relación independiente entre él y la consecuencia del estudio sin que éste sea resultado del tratamiento. Además un confusor no puede ser un intermediario entre el tratamiento y la consecuencia. Por ejemplo, la relación entre la dieta y la enfermedad coronaria puede explicarse observando cuanto colesterol tiene el paciente, pero el colesterol no es un confusor ya que es un puente causal entre la dieta y la enfermedad.

En los estudios en los que se evalúa un procedimiento quirúrgico, los confusores pueden tomar la forma de una indicación para el uso de ese procedimiento. Tal confusor por indicación puede ser extremadamente importante en estudios de eficacia o de seguridad del procedimiento. Por ejemplo, supóngase que todos los pacientes que recibieron el tratamiento A están más enfermos que los que recibieron el tratamiento B, y que

se encontró una diferencia estadísticamente significativa mostrando que el tratamiento B es el mejor. Hasta aquí no es válido concluir que B es verdaderamente el mejor tratamiento. Dado que la severidad de la enfermedad está asociada al resultado y que la severidad también se asocia con la elección del tratamiento, los efectos de éste no se pueden separar de los efectos de la severidad de la enfermedad. La única manera de comparar la efectividad de los dos tratamientos es asegurar que el diseño del estudio incluya pacientes con el mismo grado de la enfermedad en ambos tratamientos, y que la elección de estos no dependa del grado de la enfermedad.

2.8. Análisis de Supervivencia

El Análisis de Supervivencia es un conjunto de técnicas estadísticas que se usa para analizar datos con el interés en el tiempo de ocurrencia de un evento. Ejemplos de respuestas de interés incluyen el tiempo hasta que falla un componente de una maquina, el tiempo hasta el fallo de una lampara o el tiempo hasta un embarazo. Al tipo de evento se le suele llamar tipo de fallo o fracaso y al tiempo hasta la ocurrencia del fracaso es llamado tiempo de fallo o tiempo de supervivencia. Para determinar un tiempo de fallo preciso se requiere que el inicio del estudio, la escala para medir el tiempo y el tipo de fallo se defina sin ambigüedad.

2.8.1. Características de los datos de supervivencia

Una característica especial en este tipo de datos es cuando el objeto de estudio no finaliza con el seguimiento, en este caso se dice que el tiempo de supervivencia es censurado. También se considera como censurado cuando otro tipo de fallo interviene o cuando finaliza el estudio y el individuo no presenta el evento de interés. Por ejemplo, en un estudio clínico donde se están analizando dos tipos de tratamientos y el tipo de fallo es la muerte por la enfermedad, un paciente abandona el estudio debido a un cambio de residencia o muere por causas ajenas a este malestar, entonces esto no nos permite saber el tiempo de supervivencia dando así un tiempo de fallo censurado. A los tipos de censura anteriores se les denomina datos censurados a la derecha debido a que se sabe que el tiempo de supervivencia es mayor al tiempo de censura. También existen casos que el tiempo de censura es a la izquierda, es decir, el fracaso ya ocurrió cuando se observa por primera vez y censurado por intervalo que es cuando se tiene una cota superior e inferior de cuando pudo haber ocurrido el fracaso. Este trabajo sólo se enfocará en los datos censurados por la derecha.

En la figura 2.1 se ilustra un estudio con duración de 10 años, los tiempos censurados por abandono o perdida de seguimiento se muestran con un circulo negro mientras que el círculo blanco los que fueron censurados debido a que no presentaron el fracaso a lo largo del seguimiento y los casos que mostraron el fracaso en un tiempo determinado a lo largo del estudio se muestran con un rombo.

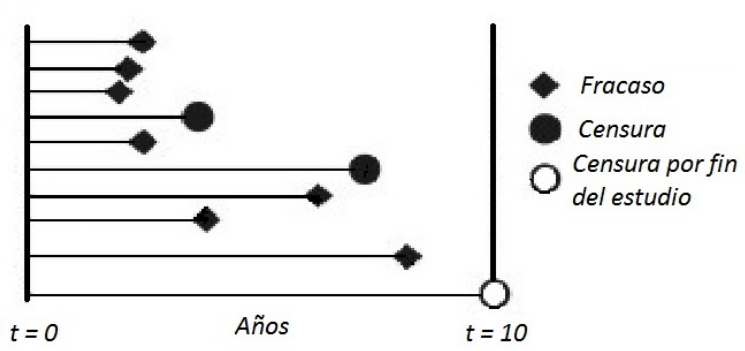


Figura 2.1: Ilustración de tiempos de supervivencia y censurados

Otra característica es que los tiempos de supervivencia suelen distribuirse de forma asimétrica, entonces no es factible suponer que los datos se distribuyan de forma normal. Una manera de solucionar esto es buscando una transformación para hacer los datos más simétricos o adoptar distribuciones de probabilidad con esta característica.

2.8.2. Análisis de supervivencia univariado

El análisis de supervivencia univariado es el análisis de datos de supervivencia para casos expuestos a un sólo tipo de fallo. Para empezar con este análisis se supone que la población es homogénea (todos son susceptibles a experimentar el evento de interés) y que cada individuo tiene un tiempo de supervivencia, es decir, no hay tiempos censurados. El tiempo de supervivencia de un individuo t puede ser considerado como el valor de una variable aleatoria T que puede tomar valores no negativos y tiene una distribución de probabilidad, entonces se denotará a la función de densidad de probabilidad por $f(t)$ y a la función de distribución que representa la probabilidad que el tiempo de supervivencia sea menor al tiempo t es definida por:

$$F(t) = P\{T < t\} = \int_0^t f(u)du$$

Es de interés representar la probabilidad que un individuo tenga un tiempo de supervivencia mayor al tiempo t, para hacerlo existe una función muy importante que es la función de supervivencia y que es dada por:

$$S(t) = Pr\{T \geq t\} \tag{2.9}$$

Otra función muy importante es la fuerza de mortalidad, esta nos da una idea de como va cambiando la probabilidad de presentar el evento de interés en el siguiente instante de tiempo. Se define como la probabilidad que un individuo muera en el tiempo t dado que ha sobrevivido hasta ese tiempo. Para definir la función de mortalidad, sea δt un tiempo infinitesimal y considere la probabilidad condicional que un individuo

experimente el tipo de fallo en el intervalo de tiempo $[t, t+\delta t)$ dado que ha sobrevivido hasta el tiempo t, esto puede expresarse como:

$$h(t) = \lim_{\delta t \to 0} \left[\frac{P[t \leq T < t + \delta t | T \geq t]}{\delta t} \right]$$

Una relación importante de la fuerza de mortalidad con la función de supervivencia y la función de densidad se obtiene al aplicar las definiciones de derivada y de probabilidad condicionada como se puede observar en la siguiente ecuación:

$$h(t) = \lim_{\delta t \to 0} \left[\frac{\frac{P[t \leq T < t+\delta t]}{P[T \geq t]}}{\delta t} \right] = \frac{1}{S(t)} \lim_{\delta t \to 0} \left[\frac{F(t+\delta t) - F(t)}{\delta t} \right] = \frac{f(t)}{S(t)} \tag{2.10}$$

De lo observado anteriormente, la fuerza de mortalidad se puede expresar en términos de la función de supervivencia por medio de la siguiente relación:

$$h(t) = -\frac{d}{dt} \log(S(t)) \tag{2.11}$$

de manera análoga, una expresión de la función de supervivencia en términos de la fuerza de mortalidad es:

$$S(t) = \exp(-H(t)) \tag{2.12}$$

donde $H(t)$ es la fuerza de mortalidad integrada y se define por:

$$H(t) = \int_0^t h(u) du \tag{2.13}$$

2.8.3. Estimación no paramétrica de la función de supervivencia

Un método usado para describir los datos de supervivencia es el estimador del producto límite de la función de supervivencia propuesto por Kaplan-Meier. Supóngase que hay n individuos con tiempos de supervivencia observada $t_1, t_2, \ldots, t_n$, algunas de estas observaciones podría ser censurada a la derecha y podría ser que mas de un individuo tenga el mismo tiempo de supervivencia, por lo tanto supóngase que hay r tiempos de supervivencia entre los individuos, donde $r \leq n$ y se ordenan los tiempos de muerte ascendentemente $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$. El número de individuos los cuales están vivos justo antes del tiempo $t_{(j)}$ incluyendo a los que mueren en ese tiempo, se denotará por n_j, para $j = 1, 2, \ldots, r$, y d_j denotará el numero de muertes en este tiempo. El intervalo de tiempo $(t_{(j)} - \delta, t_{(j)}]$, donde δ es un número infinitesimal, incluye un tiempo de muerte. Ya que hay n_j individuos vivos justo antes de $t_{(j)}$ y d_j muertos en $t_{(j)}$, la probabilidad que un individuo muera durante el intervalo $(t_{(j)} - \delta, t_{(j)}]$ es estimado por d_j/n_j. La probabilidad correspondiente estimada de supervivencia en este intervalo es $(n_j - d_j)/n_j$. Si un tiempo de supervivencia y un tiempo de censura ocurren simultáneamente, el

tiempo de supervivencia del tiempo de censura se considera que ocurre inmediatamente después del tiempo de supervivencia cuando se calculan los valores de n_j.

El intervalo $(t_{(j)}, t_{(j+1)} - \delta]$, el tiempo inmediatamente antes del siguiente tiempo de muerte no contiene muertes, por lo tanto, la probabilidad de sobrevivir en este intervalo es uno. Entonces la probabilidad conjunta de sobrevivencia de $(t_{(j)} - \delta, t_{(j)}]$ y $(t_{(j)}, t_{(j+1)} - \delta]$ puede ser estimada por $(n_j - d_j)/n_j$. En el límite, cuando δ tiende a cero, $(n_j - d_j)/n_j$ se convierte en un estimador de la probabilidad de supervivencia de $(t_{(j)}, t_{(j+1)}]$.

Supóngase que las muertes de los individuos en la muestra ocurren independientemente. Entonces, la función de supervivencia estimada en el k-ésimo intervalo de tiempo construido de $t_{(k)}$ a $t_{(k+1)}$, $k = 1, 2, \ldots, r$ donde $t_{(r+1)}$ es ∞, se calcula como la probabilidad de supervivencia después de t_k, esta es la probabilidad de sobrevivir en el intervalo $t_{(k)}$ a $t_{(k+1)}$ y todos los intervalos anteriores. A este estimador se le conoce como el estimador de Kaplan-Meier de la función de supervivencia, el cual está dado por:

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right), \quad t \in [t_{(k)}, t_{(k+1)}) \tag{2.14}$$

para $k = 1, \ldots, r$, con $\hat{S}(t) = 1$ para $t < t_{(1)}$. Si $t_{(r)}$ es censurado, entonces $\hat{S}(t)$ no esta definido para $t > t_{(r)}$, por otro lado, si es una observación sin censura, $n_r = d_r$, y así $\hat{S}(t)$ es cero para $t \geq t_{(r)}$. Una gráfica del estimador de Kaplan-Meier de la función de supervivencia es una función escalonada en la cual la estimación de probabilidades de supervivencia son constantes entre tiempos de muerte adyacentes y decreciente en cada tiempo de muerte.

Un estimador de la fuerza de mortalidad integrada puede ser obtenida de la ecuación (2.12):

$$\hat{H}(t) = -\log\{\hat{S}(t)\} = -\sum_{j=1}^{k} \log \left(\frac{n_j - d_j}{n_j} \right) \tag{2.15}$$

Aproximando por expansión de series de Taylor:

$$\log \left(\frac{n_j - d_j}{n_j} \right) = \log \left(1 - \frac{d_j}{n_j} \right) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \left(-\frac{d_j}{n_j} \right)^n \approx -\frac{d_j}{n_j}$$

De lo anterior se deduce otra forma de estimar la fuerza de mortalidad integrada conocida como el estimador de Nelson-Aalen es:

$$\hat{H}(t) = \sum_{t_j \leq t} \frac{d_j}{n_j} \tag{2.16}$$

Para muestras de tamaño grandes los dos estimadores son similares.

2.8.4. Distribuciones de Modelos Paramétricos

Esta sección se enfocará en la función de distribución Weibull que es utilizada en el modelo que se ocupará en esta aplicación, también se presenta la distribución exponencial para tener un mayor entendimiento de la distribución Weibull.

Distribución Exponencial

La manera más sencilla para suponer la distribución de la fuerza de mortalidad es que es constante sobre el tiempo, es decir,

$$h(t) = \theta$$

con $0 \leq t < \infty$ y $\theta > 0$. Por la ecuación (2.12) se obtiene:

$$S(t) = \exp\left(-\int_0^t \theta \, du\right) = \exp(-\theta t)$$

Entonces T es una variable aleatoria que se distribuye exponencialmente con parámetro θ y que será denotado por $T \sim \text{Exp}(\theta)$, cuando $\theta = 1$ entonces se dice que la variable aleatoría tiene una distribución aleatoria exponencial estándar. Por lo anterior y la ecuación 2.10 la función de densidad esta dada por:

$$f(t) = \theta \exp(-\theta t)$$

Distribución Weibull

La función de distribución Weibull fue propuesta por W. Weibull en 1951 en pruebas de fiabilidad en la industria, debido a la forma que puede tomar esta función juega un papel central en el análisis de supervivencia. Sea Y una variable aleatoria exponencial estándar, definimos:

$$X = \lambda Y^{1/\gamma}, \quad \lambda > 0, \quad \gamma > 0$$

Entonces se dice que X tiene una distribución Weibull (denotado por $X \sim W(\gamma, \lambda)$) con parámetro de forma γ y parámetro de escala λ. Esta distribución tiene una función de fuerza de mortalidad definida por:

$$h(x) = \frac{\gamma x^{\gamma-1}}{\lambda^\gamma}$$

con $0 \leq x < \infty$. La función $h(x)$ es monótona creciente si $\gamma > 1$, monótona decreciente si $\gamma < 1$, y para $\gamma = 1$ es la distribución exponencial con parámetro λ. Por la ecuación (2.12) la función de supervivencia esta dada por:

$$S(x) = \exp\left[-\left(\frac{x}{\lambda}\right)^\gamma\right]$$

y por la ecuación 2.10 se deduce que la función de densidad es:

$$f(x) = \frac{\gamma}{\lambda}\left(\frac{x}{\lambda}\right)^{\gamma-1}\exp\left[-\left(\frac{x}{\lambda}\right)^{\gamma}\right]$$

Esta distribución tiene una cola del lado derecha más larga que del lado izquierdo, entonces esta función es sesgada del lado derecho, lo cual la hace ideal para modelar datos de supervivencia.

2.8.5. Información Concomitante

En datos de supervivencia se llega a registra información adicional de la que se cree que depende el tiempo de supervivencia, a las variables donde se observa este tipo de información se les llama variables explicativas. Todas estas variables explicativas se pueden considerar como un vector de covariables $\boldsymbol{Z}^T = (Z_1, Z_2, \ldots, Z_p)$ donde Z_i denota alguna característica del individuo. Se les puede clasificar según su escala de medición como:

- Variables Cuantitativas: Son variables que son caracterizadas por alguna información numérica que se le puede asociar a los individuos de una población, por lo tanto, se pueden realizar operaciones aritméticas con ella. Se clasifican en continuas, cuando toman cualquier valor dentro de un intervalo numérico y discretas cuando solamente toman valores específicos y no toma valores intermedios a estos.
- Variables Cualitativas: Son variables que están relacionadas con características que poseen los individuos de la población o cosas no son medibles. Cada una de esas cualidades presentes se denomina atributo o categoría. A este tipo de variables se les llama factor y a sus categorías se les denomina niveles para el factor. En este tipo de variables se utilizan dos escalas, la escala ordinal, estas presentan un orden en sus categorías; y la escala nominal, a diferencia de las anteriores no se les puede asignar un orden.

2.8.6. Modelo de riesgos proporcionales

Un modelo para los datos de supervivencia es el modelo de riesgos proporcionales. Este modelo fue propuesto por Cox en 1972, también se le conoce como modelo de regresión de Cox. El modelo se basa en la suposición que la tasa de muerte de un individuo en algún tiempo en un grupo de estudio es proporcional a la tasa de muerte a algún tiempo de un individuo similar en otro grupo.

Para describir el modelo supóngase que la fuerza de mortalidad depende de las variables explicativas $Z_1, Z_2, \ldots, Z_p$ y sean $z_1, z_2, \ldots, z_p$ sus valores. El conjunto de valores de las variables explicativas será denotado por $\mathbf{z} = (z_1, z_2, \ldots, z_p)^T$. Sea $h_0(t)$ la fuerza de mortalidad para un individuo con características $\mathbf{z} = \mathbf{0}$, la cual es llamada la fuerza de mortalidad de base. Sea $h(\boldsymbol{z})$ una función de los valores del vector $\mathbf{z}$, entonces la fuerza de mortalidad para el i-ésimo individuo se puede escribir como:

$$h_i(t) = h(\mathbf{z}_i)h_0(t)$$

La función $h(\mathbf{z})$ se puede interpretar como la fuerza de mortalidad en el tiempo t para un individuo cuyo vector de variables explicativas es $\mathbf{z}_i$ relativo a la fuerza de mortalidad del individuo con características $\mathbf{z} = \mathbf{0}$. Ya que esta función no puede ser negativa es conveniente que $h(\mathbf{z}_i) = \exp(\gamma_i)$ donde $\gamma_i = \boldsymbol{\eta}^T\mathbf{z}_i$ con $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_p)^T$ es el vector de coeficientes de las variables explicativas $\mathbf{z}$ en el modelo. El valor de γ_i es llamado la componente lineal del modelo, aunque también es conocido como el índice de pronósticos para el i-ésimo individuo. El modelo de riesgos proporcionales es $h_i(t) = \exp(\boldsymbol{\eta}^T\mathbf{z}_i)h_0(t)$, que puede ser escrita como:

$$\log\left\{\frac{h_i(t)}{h_0(t)}\right\} = \exp(\boldsymbol{\eta}^T\mathbf{z}_i)$$

La ecuación anterior podría ser considerada como un modelo lineal para el logaritmo de la razón de fuerzas de mortalidad.

2.8.7. Decremento Múltiple

Hay situaciones donde no es apropiado aplicar métodos de supervivencia univariados, tales casos es donde hay varios tipos de fallo. En general, situaciones donde un individuo puede experimentar más de un tipo de evento y la ocurrencia de este oculta la ocurrencia de otros tipos de evento. Por ejemplo, supóngase que se da seguimiento a un grupo de pacientes diagnosticados con cáncer de mama para observar su muerte por esta enfermedad. Las técnicas de supervivencia pueden ser usadas si todos los pacientes murieron por la enfermedad o sobrevivieron. Sin embargo, en datos reales se da el caso en que los individuos mueren de otras causas antes de morir por el cáncer, esta es una situación de decremento múltiple ya que la muerte de otras causas impide la ocurrencia del evento de interés. En esta sección se presenta una introducción al decremento múltiple.

Para desarrollar una metodología para este tipo de datos se supone que se tiene una población de individuos sujeta a J tipos de fallo los cuales están operando simultáneamente entre ellos, además, el mecanismo de censura es por la derecha y es independiente del tiempo de supervivencia. También se hacen las siguientes suposiciones:

- Si algún tipo de fallo ocurre, este puede ser por uno de los J distintos tipos de fallo.
- Cada individuo es susceptible a sufrir cualquier tipo de fallo.

Se supone que a cada causa de fallo j esta asociado con una variable aleatoria T_j continua que es el tiempo de supervivencia en condiciones hipotéticas donde cada causa de fallo es único. De esta forma cada individuo está asociado a un vector de variables

aleatorias $\boldsymbol{T} = (T_1, T_2, \ldots, T_J)^T$. Se define la función de supervivencia multivariada de los tiempos de fallo $T_1, T_2, \ldots, T_J$ como:

$$S_{1,\ldots,J}(t_1, t_2, \ldots, t_J) = Pr\left\{\bigcap_{i=1}^{J}(T_i > t_i)\right\} \tag{2.17}$$

para todo $t_j > 0$. La función de distribución acumulada conjunta correspondiente a $S_{1,\ldots,J}(\cdot)$ se define por $F_{t_1,\ldots,t_J}$. Para la fuerza de mortalidad multivariada con respecto a t_j en $\boldsymbol{t} = (t_1, t_2, \ldots, t_J)$ se define por:

$$h_j(\boldsymbol{t}) = \lim_{h\to 0}\frac{1}{h}Pr\left\{(t < T_j \leq t + h)\bigcap_{\substack{i=1\\ i\neq j}}^{J}(T_i > t)|\bigcap_{i=1}^{J}(T_i > t)\right\} = -\frac{\partial \log S_{1,\ldots,J}(\boldsymbol{t})}{\partial t_i}$$

donde la última igualdad es obtenida por la definición de probabilidad condicional y la regla de la cadena de las derivadas. Ya que todos los tipos de fallo están actuando simultáneamente y por hipótesis un sólo tipo de fallo ocurre, no es posible observar $T_1, T_2, \ldots, T_J$ en conjunto, en vez de hacer eso, cada individuo esta caracterizado por el par aleatorio observado (T, C) donde $T = \text{mín}\{T_1, T_2, \ldots, T_J\}$ que se define como el tiempo de supervivencia real y $C = \{1, \ldots, J\}$ es el índice de tipo de fallo. Entonces se tiene una distribución bivariada con T continua y C discreta. Para continuar con la metodología se definen dos tipos de probabilidades:

- Probabilidades netas: Con excepción de una causa de fallo j-ésima, todas los otros tipos de fallo son eliminados.
- Probabilidades crudas: Todas los tipos de fallo están presentes.

La probabilidad de sobrevivir de cualquier causa hasta el tiempo t se define por la función de supervivencia global, que se expresa como:

$$S_T(t) = Pr\{T > t\} = Pr\left\{\bigcap_{j=1}^{J}(T_j > t)\right\} = S_{1,\ldots,J}(t, t, \ldots, t) \tag{2.18}$$

Ya que (2.18) es una función de una sola variable, por la ecuación (2.11) se deduce que la fuerza de mortalidad global es:

$$h_T(t) = \lim_{h\to 0}\frac{1}{h}Pr\{(t < T_j \leq t + h)|T > t\} = -\frac{d\log S_T(t)}{dt} = -\frac{d\log S_{1,\ldots,J}(t, t, \ldots, t)}{dt}$$

y por la ecuación (2.12) se produce la relación inversa:

$$S_T(t) = \exp\left[-\int_0^t h(u)du\right] = \exp[-H_T(t)]$$

donde $H_T(t)$ es la función de riesgo acumulada global. Se define la fuerza de mortalidad cruda la cual describe la tasa instantánea de morir debido a la causa j al tiempo t cuando todas las causas actúan simultáneamente condicionada a que no ha ocurrido ningún tipo de fallo por:

$$\begin{aligned} h_j(t) &= \lim_{h\to 0}\frac{1}{h}Pr\left\{(t<T_j\leq t+h)\bigcap_{\substack{i=1\\ i\neq j}}^{J}(T_i>t)\Big|\bigcap_{i=1}^{J}(T_i>t)\right\} \\ &= -\frac{\partial \log S_{1,\ldots,J}(t_1,t_2,\ldots,t_J)}{\partial t_j}\bigg|_{t_i=t} \end{aligned}$$

con $j=1,\ldots,J$ y $t>0$. Para dar la relación que tiene la fuerza de mortalidad global $h_T(t)$ con la fuerza de mortalidad cruda $h_j(t)$ se observa que aplicando la regla de la cadena a la función $\log S_{1,\ldots,J}(t,\ldots,t)$ se obtiene el siguiente resultado:

$$\frac{d\log S_{1,\ldots,J}(t,\ldots,t)}{dt}=\sum_{i=1}^{J}\frac{\partial \log S_{1,\ldots,J}(t_1,\ldots,t_J)}{\partial t_i}\bigg|_{t_i=t}$$

con lo anterior se deduce la propiedad aditiva de las fuerzas de mortalidad crudas $h_T(t)=h_1(t)+\cdots+h_J(t)$. La probabilidad condicional de ocurrir el tipo de fallo j en el intervalo $(t,t+h)$ dado que es libre de cualquier otro evento es aproximadamente $h_j(t)h$. La probabilidad no condicionada de que el tipo de fallo sea j en el intervalo de tiempo $(t,t+h)$ es aproximadamente $h_j(t)S_T(t)h$, así la distribución de la probabilidad cruda del tiempo de muerte por la causa j es:

$$Q_j(t)=Pr\{T\leq t,C=j\}=\int_0^t h_j(u)S_T(u)du \tag{2.19}$$

y la probabilidad cruda que el tipo de fallo j de un individuo se dé después del tiempo t es definida por:

$$P_j(t)=Pr\{T>t,C=j\}=\int_t^\infty h_j(u)S_T(u)du$$

La proporción de muertes esperadas por la causa j se define por:

$$p_j=Pr\{C=j\}=\int_0^\infty h_j(u)S_T(u)du=Q_j(\infty)=P_j(0) \tag{2.20}$$

con $\sum_{j=1}^{J}p_j=1$. Se tiene que para cualquier función multivariada de supervivencia multivariada cumple que [3]:

$$\frac{dP_j(t)}{dt}=-\frac{\partial \log S_{1,\ldots,J}(t_1,t_2,\ldots,t_J)}{\partial t_i}\bigg|_{t_i=t}$$

Así la función de densidad cruda la cual representa el riesgo que un individuo experimente un tipo de fallo al tiempo t de la causa j asociada con la función de supervivencia cruda $P_j(t)$ puede ser calculada como:

$$f_j(t) = -\frac{\partial \log S_{1,\ldots,J}(t_1, t_2, \ldots, t_J)}{\partial t_i}\bigg|_{t_i=t}$$

De lo anterior se deduce la siguiente relación:

$$h_j(t) = \frac{f_j(t)}{S_T(t)}$$

Por lo tanto la función de densidad global se obtiene por:

$$f_T(t) = \sum_{j=1}^{J} f_j(t) = S_T(t)h_T(t) = -\frac{dS_T(t)}{dt}$$

Entonces la función de distribución acumulada y de supervivencia global están relacionadas con las funciones $Q_j(t)$ y $P_j(t)$ por medio de las siguientes igualdades:

$$F_T(t) = 1 - S_T(t) = \sum_{j=1}^{J} Q_j(t)$$

$$S_T(t) = \sum_{j=1}^{J} F_j(t)$$

y por definición se obtiene que $F_T(t) + S_T(t) = 1$.De lo anterior se deduce que las funciones $Q_j(t)$ y $P_j(t)$ no son distribuciones propias ya que $P_j(t) + Q_j(t) = p_j \leq 1$, no obstante, se pueden definir distribuciones condicionales propias en términos de esas distribuciones. Entonces, la función de distribución propia es definida por:

$$S_j^*(t) = Pr(T > t|C = j) = \frac{P_j(t)}{P_j(0)} = \frac{1}{p_j}P_j(t) = \frac{1}{p_j}\int_t^\infty h_j(u)S_T(u)du$$

con $j = 1, 2, \ldots, J$. Entonces la probabilidad condicional de presentar un tipo de fallo antes del tiempo t y en presencia de todas las causas dado que el individuo tendrá una falla por la causa j se define por $F_j^*(t) = 1 - S_j^*(t) = Pr(T \leq t|C = j)$. Derivando $F_j^*(t)$ se deduce la función de densidad:

$$f_j^*(t) = -\frac{1}{p_j}\frac{dP_j(t)}{dt} = -\frac{dS_j^*(t)}{dt} = \frac{1}{p_j}h_j(t)S_T(t)$$

La fuerza de mortalidad de $S_j^*(t)$ es:

$$h_j^*(t) = -\frac{d\log S_j^*(t)}{dt} = \frac{f_j^*(t)}{S_j^*(t)} = \frac{h_j(t)S_T(t)}{\displaystyle\int_t^\infty h_j(u)S_T(u)du} = -\frac{1}{P_j(t)}\frac{dP_j(t)}{dt} \tag{2.21}$$

Ahora se observa que:

$$h_j(t) = -\frac{1}{S_T(t)}\frac{dP_j(t)}{dt} = \frac{dP_j(t)/dt}{\sum_{i=1}^{J} P_i(t)} \tag{2.22}$$

Tomando la razón de (2.22) y (2.21) se tiene la probabilidad condicional de que un individuo falle por la causa j después del tiempo t, dado que ha sobrevivido a todas las causas hasta ese tiempo, que se expresa por:

$$\frac{h_j(t)}{h_j^*(t)} = \frac{P_j(t)}{S_T(t)} = \frac{P_j(t)}{\sum_{i=1}^{J} P_i(t)} = Pr(C = j|T > t) = p_j(t)$$

Estimación no paramétrica

En esta sección se generalizan técnicas de estimación no paramétrica como el estimador de Kaplan-Meier y Nelson Aalen para datos de decremento múltiple. Considérese el conjunto de datos independientes censurados por la derecha de un modelo de decremento múltiple homogéneo, y supóngase que se tienen r distintos tiempos de supervivencia ordenados de todos los distintos tipos de fallo $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$. Sea d_{ji} el número de fallos del tipo j al tiempo t_i y n_i el número de individuos en riesgo al tiempo t_i, $i = 1, 2, \ldots, r$ y $j = 1, 2, \ldots, J$. Entonces la función de verosimilitud se puede expresar como [4]:

$$L = \prod_{i=1}^{n}\left[\prod_{j=1}^{J}(h_{ij})^{d_{ij}}(1-h_i)^{n_i-d_i}\right]$$

donde $d_i = \sum_{j=1}^{J} d_{ij}$ es el número de fallos al tiempo t_i y $h_i = h_T(t_i) = \sum_{j=1}^{J} h_{ij}$ es la fuerza de mortalidad global al tiempo t_i. El estimador de máxima verosimilitud de h_{ij} se deduce al maximizar la función de verosimilitud, siendo $\hat{h}_{ij} = d_{ij}/n_i$. Por lo tanto, el estimador de Nelson-Aalen de la fuerza de mortalidad acumulada para la falla j es:

$$\hat{H}_j(t) = \sum_{t_i \le t}\frac{d_{ij}}{n_i}$$

con $t > 0$ y $j = 1, \ldots, J$. Entonces, el estimador de Nelson-Aalen para la fuerza de mortalidad acumulada global es:

$$\hat{H}_T(t) = \sum_{j=1}^{J}\hat{H}_j(t)$$

De esta manera el estimador de Nelson-Aalen para la función de supervivencia global se expresa como:

$$\hat{S}_T(t) = \exp -\hat{H}_T(t) = \exp\left[-\sum_{j=1}^{J} \hat{H}_j(t)\right]$$

Por lo tanto un estimador para la función $Q_j(t)$ por (2.19) es:

$$\hat{Q}_j(t) = \sum_{t_i \leq t} \hat{S}_T(t_i^-)\frac{d_{ij}}{n_i}$$

donde $\hat{S}_T(t_i^-) = \lim_{\varepsilon \to 0} \hat{S}_T(t_i - \varepsilon) = \hat{S}_T(t_{i-1})$. Sea $t_{(r)}$ es el último tiempo de supervivencia de una observación no censurada. Debido a (2.20) la probabilidad de finalmente presentar el fallo por la causa j se estima por:

$$\hat{p}_j = \hat{Q}_j(t_{(r)}) \quad j = 1, \ldots, J$$

Similarmente, el estimador de $F_j^*(t)$ se expresa por:

$$\hat{F}_j^*(t) = \frac{\hat{Q}_j(t)}{\hat{p}_j}$$

Un estimador alternativo para la función de supervivencia global es el estimador de Kaplan-Meier:

$$\hat{S}_T(t) = \prod_{t_i \leq t}[1 - \hat{h}_T(t_i)] = \prod_{t_i \leq t}\left[1 - \frac{d_i}{n_i}\right]$$

para $t > 0$ y $d_i = \sum_{j=i}^{J} d_{ij}$ es el número total de fallos de todas las causas al tiempo t.

Modelo de riesgos proporcionales

Para los modelos de regresión en el caso de decremento múltiple sea $\mathbf{z}_i$ un vector de variables explicativas para el i-ésimo individuo, para incluir variables explicativas en un modelo de regresión se hace una adaptación de la definición univariada de riesgos proporcionales a la fuerza de mortalidad cruda, esto es:

$$h_j(t; \mathbf{z}_i) = \phi_i(\mathbf{z}_i)h_{0j}(t), \quad t > 0$$

donde $h_{0j}(t)$ es la función de riesgo base ($\mathbf{z}_i = 0$) para el modo de fallo j, y $\phi_i(\mathbf{z}_i)$ es una función positiva del vector de covariables.

Modelo de Larson & Dinse (1985)

El modelo de mezclas propuesto por Larson & Dinse [5] se basa en una formulación semi-Markov descrita por [6]. Se supone que la censura es por la derecha y que el mecanismo de este tipo de censura es no-informativo en el sentido que si el tiempo de supervivencia t de un individuo está censurado al tiempo c todo lo que se sabe es que $t > c$ y el mecanismo de censura no proporciona alguna otra información sobre el tipo de eventual fallo de ese individuo [7]. Considérese un individuo expuesto a J distintos riesgos competitivos o tipos de fallo. El evento estocástico de interés es el par aleatorio (C, T), donde C toma valores del conjunto $\{0, 1, 2, \ldots, J\}$ para indicar el tipo de fallo o censura por la derecha cuando $C = 0$ y T es una variable aleatoria no negativa que representa el tiempo para el fallo. Se supone además que el mecanismo de censura es no informativo, es decir, las observaciones censuradas solo dan una cota inferior para el tiempo de fallo.

Sean $\mathbf{z}_i$ y $\mathbf{x}_i$ dos vectores de covariables p y q dimensionales para el i-ésimo individuo, con $i = 1, \ldots, n$. Los efectos de las covariables sobre las probabilidades de eventual causa de muerte específica puede ser modelada usando un modelo de regresión logística, dado por:

$$p_j = p_j(\mathbf{x}_i) = Pr\{C = j|\mathbf{x}_i\} = \frac{\exp(\delta + \boldsymbol{\delta}_j^T \mathbf{x}_i)}{\sum_{l=1}^{J} \exp(\delta + \boldsymbol{\delta}_l^T \mathbf{x}_i)}$$

donde δ_j es una constante escalar y $\boldsymbol{\delta_j}$ es un vector fila de p coeficientes de regresión. Por unicidad, definimos $\delta_J = 0$ y $\boldsymbol{\delta}_J = \mathbf{0}$. Para el efecto del vector de covariables $\mathbf{z}_i$ sobre la fuerza de mortalidad se supone un modelo de riesgos proporcionales, de esto se deduce:

$$h_j^*(t; \mathbf{z}_i) = h_{0j}(t) \exp(\boldsymbol{\beta}_j^T \mathbf{z}_i)$$

donde $1 \leq j \leq J$, $\boldsymbol{\beta}_j$ es un vector fila de q coeficientes de regresión y $h_{0j}(t)$ es la función de la fuerza de mortalidad base para la falla de tipo j. De lo anterior definimos la fuerza de mortalidad integrada como $H_{0j}(t)$, entonces la función de supervivencia para el tiempo de fallo dado el fracaso j se define como:

$$S_j^*(t; \mathbf{z}_i) = \exp\left[-H_{0j}(t) \exp(\boldsymbol{\beta}_j^T \mathbf{z}_i)\right] \tag{2.23}$$

Para definir la función de verosimilitud, sea $f_j^*(t; \mathbf{z}_i)$ la función de densidad correspondiente a la j-ésima función de supervivencia condicional, entonces el individuo i que fallece por la causa j en el tiempo t contribuye con $p_j(\mathbf{x}_i) f_j^*(t; \mathbf{z}_i)$ a la función de verosimilitud. Por otra parte un individuo que tiene una observación de censura en el tiempo t contribuye con $p_j(\mathbf{x}_i) S_j^*(t; \mathbf{z}_i)$ a la función de verosimilitud. De esta forma la función de verosimilitud correspondiente a n individuos se puede expresar como:

$$L_n(\theta) = \prod_{i=1}^{n} \left(\prod_{j=1}^{J} [p_{ij} f_j(t_i)]^{c_{ij}} \right) \left(\sum_{j=1}^{J} p_j S_j(t_i) \right)^{1-c_i}$$

donde $c_1, c_2 \ldots, c_n$ son los indicadores de censura se definen para $1 \leq i \leq n$ del siguiente modo:

$$c_i = \begin{cases} 1 & \text{si el individuo } i \text{ no está censurado} \\ 0 & \text{si el individuo } i \text{ está censurado} \end{cases}$$

Y se define c_{ij} como los indicadores de la causa de muerte:

$$c_{ij} = \begin{cases} 1 & \text{si el individuo } i \text{ murió por la causa } j \\ 0 & \text{otro caso} \end{cases}$$

para $1 \leq i \leq n$ y $1 \leq j \leq J$.

Capítulo 3

Datos faltantes

En estadística se busca explicar y entender fenómenos de interés por medio de recolección de datos. Los datos faltantes en estadística significa que hay información que por alguna razón no esta disponible en la base de datos que se desea analizar y esto conlleva a afectar la habilidad para inferir o deducir la naturaleza del fenómeno de interés, por esto es necesario tratar el conjunto de datos faltantes.

En estudios epidemiológicos suelen haber una gran cantidad de datos faltantes, ya que puede existir la ausencia de participación del paciente. En este capitulo se dan conceptos relacionados con datos faltantes, como los mecanismos que dejan a los datos faltantes y algunas formas que para tratarlos para su análisis. Entre estos métodos se esta interesado en el de imputación múltiple MICE que se ocupará en la aplicación de este trabajo.

3.1. Notación

Antes de empezar con la teoría de datos faltantes, se va a dar la notación utilizada en este capitulo. Supóngase que se tiene una muestra de tamaño n y definase por $\boldsymbol{Y} = (Y_{i,1}, Y_{i,2}, \ldots, Y_{i,p})$ las p variables que se recolectaron del $i-$ésimo individuo, $i = 1, \ldots, n$ y $\boldsymbol{Y}$ la matriz de n por p que contiene los datos. Denotese por $\mathbf{Y}_{i,O}$ el subconjunto de datos que son observados y $\mathbf{Y}_{i,F}$ a el subconjunto de datos que son faltantes para cada individuo $i = 1, \ldots, n$. Para cada individuo $i = 1, \ldots, n$ y para variable $j = 1, \ldots, p$, sea M la matriz indicadora de respuesta que se define por $M_{i,j} = 1$ si $Y_{i,j}$ es observado y $M_{i,j} = 0$ si $Y_{i,j}$ es faltante.

3.2. Mecanismos que conducen a los datos faltantes

El sistema de clasificación más usado para datos faltantes fue el que propuso Donald Rubín en 1976. Rubín especificó tres tipos distintos de datos faltantes: missing

completely at random (MCAR), missing at random (MAR) y missing not at random (MNAR). A continuación se presentan las definiciones de estas tres clasificaciones:

Definición 1 (Missing Completely at Random) *Se dice que los datos son Missing Completely at Random (MCAR) si la probabilidad que un valor sea perdido no esta relacionado a los datos observados y a los datos no observados. En notación matemática se tiene:*

$$Pr\{\mathbf{M}_i|\mathbf{Y}_i\} = Pr\{\mathbf{M}_i\} \tag{3.1}$$

Es decir, los eventos que dejan a cualquier dato particular como faltante son independientes de las variables observadas y de las variables no observadas, estos eventos ocurren de manera totalmente aleatoria.

Definición 2 (Missing At Random) *Se dice que los datos son Missing At Random (MAR) si dado el conjunto de datos observado, la distribución de probabilidad de* $\mathbf{M}_i$ *es independiente de los datos no observados, matemáticamente se expresa como:*

$$Pr\{\mathbf{M}_i|\mathbf{Y}_i\} = Pr\{\mathbf{M}_i|\mathbf{Y}_{i,O}\} \tag{3.2}$$

Es decir, MAR ocurre cuando el mecanismo de datos faltantes no es aleatorio y estos datos faltantes pueden ser producidos por las variables donde hay información completa. MAR es una suposición que no es posible verificar estadísticamente, ya que para hacerlo se tendría que tener los datos faltantes.

Definición 3 (Missing Not At Random) *Se dice que los datos son MNAR si el mecanismo que causa los datos perdidos no es MCAR ni MAR. Sobre un mecanismo MNAR, la probabilidad de que una observación sea perdida depende sobre un valor subyacente, y esta dependencia permanece aún dados los valores observados. Matemáticamente:*

$$Pr\{\mathbf{M}_i|\mathbf{Y}_i\} \neq Pr\{\mathbf{M}_i|\mathbf{Y}_{i,O}\} \tag{3.3}$$

Es decir, la probabilidad que un datos sea faltante depende de ellos mismos. Por ejemplo, una persona con un ingreso alto mensual se niega a responder esta pregunta, entonces este dato faltante depende de la variable de ingreso y no de otras variables de estudio o de la aleatoriedad.

3.3. Métodos para tratar a los datos faltantes

3.3.1. Análisis de datos completos

La ventaja de este método es la facilidad en que puede ser implementado ya que el procedimiento para tratar los datos faltantes es eliminar todos los casos con uno o más

valores faltantes sobre las variables y así realizar el análisis estadístico. Una desventaja es que podría desechar mucha de la información disponible ya que es muy común en las aplicaciones reales que más de la mitad de la muestra sea perdida y una submuestra podría seriamente degradar la detección de efectos de interés.

Si los datos son MCAR produce estimadores de medias y varianzas insesgados, mientras que para datos que son MAR y MNAR este método puede sesgar los estimadores de la media, los coeficientes de regresión y las correlaciones [8].

3.3.2. Análisis de casos disponibles

La idea sobre este método es usar toda la información disponible y producir estimadores consistentes. Por ejemplo, si se tiene dos variables continuas Y_k y Y_j entonces sus medias son tomadas sobre $(Y_k)_O$ y $(Y_j)_O$ y la covarianza de estas dos variables sobre la intersección de $(Y_k)_O$ y $(Y_j)_O$.

La desventaja de este método es que los estimadores pueden ser sesgados si los datos no son MCAR, además de haber problemas computacionales como que la matriz de correlación no sea definida positiva, correlaciones fuera del rango [-1,1] y se pueden volver más graves cuando las variables son altamente correlacionadas [9].

3.3.3. Imputación simple

El método de imputación simple rellena con un valor a cada uno de los datos faltantes creando así una base de datos completa. Uno de los métodos más sencillos de imputación simple es el de imputación de la media que fue propuesto por Wilks en 1932. Este método genera el valor de la media para datos cuantitativos y el de la moda para datos cualitativos para imputarlo en todos los datos faltantes. La desventaja de este método es que podría distorcionar la distribución, disminuir las varianzas, alterar la relación entre variables y sesgar otro estimador diferente al imputado cuando los datos no son MCAR.

Otro método es el de imputación por regresión, el cual busca el conocimiento de otras variables para realizar las imputaciones. Para realizar este método primero se construye un modelo de los datos observados tomando como predictores los casos incompletos y así reemplazarlos para los datos faltantes. La desventaja de este método es que es muy poco probable que los datos provengan del modelo de regresión propuesto; también podría incrementar la correlación de los datos debido a que los valores imputados podría estar muy correlacionados.

Un método de imputación que también utiliza ecuaciones de regresión para predecir las variables incompletas a partir de las variables completas es el de imputación por regresión estocástica, pero la diferencia es que aumenta a cada predicción un término residual que tiene una distribución, generalmente es una distribución normal. Al aumentar este residuo, reestablece la pérdida de variabilidad de los datos y elimina el sesgo asociado. Con este método de imputación se obtienen estimaciones insesgadas de los parámetros bajo datos MAR.

3.3.4. Imputación múltiple

Otro conjunto de técnicas estadísticas para tratar a los datos faltantes es el método de imputación múltiple, el cual consiste en realizar varias imputaciones de cada uno de los datos faltantes para después analizar los conjuntos de datos completos resultantes para así obtener una estimación final. Las etapas que son realizadas para estos métodos son:

- Etapa de imputación: Se crean M copias de los datos faltantes generando así M conjunto de datos.
- Etapa de análisis: Se analizan los M conjuntos de datos creando así M estimaciones de parámetros y su varianza.
- Etapa de combinación: Las estimaciones y su varianza se combinan para así tener un sólo conjunto de estimadores con su respectiva varianza.

En la figura 3.3.4 ilustra las etapas anteriores, en el primer circulo se muestra el conjunto de datos original incompleto, en la segunda etapa se muestra la imputación de los datos generando así M bases de datos completas. El paso siguiente es hacer el análisis estadístico de estas bases de datos generando así M conjuntos de resultados, y por último se aplica la etapa de combinación para tener los estimadores finales.

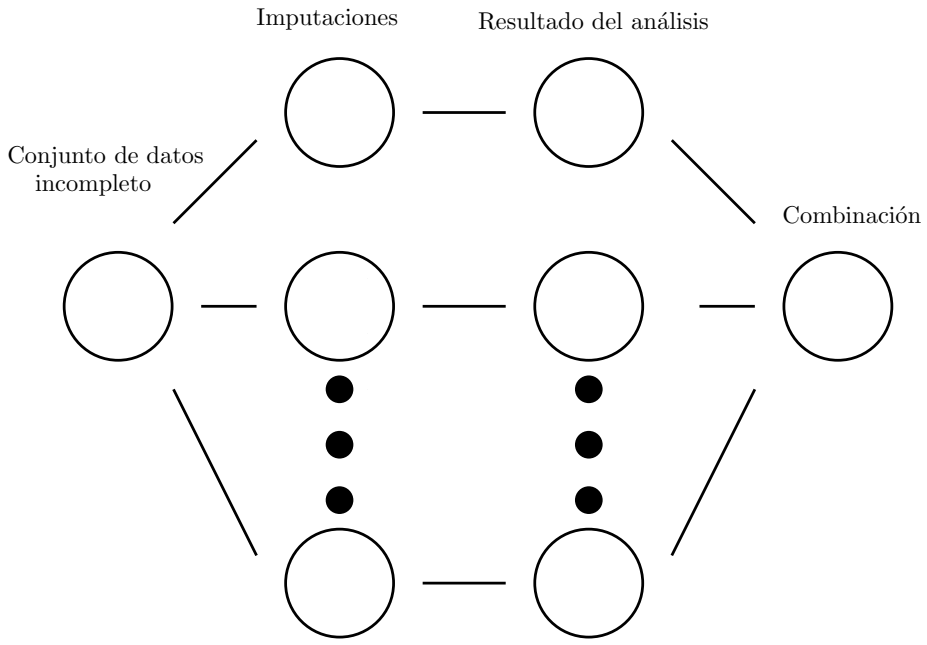


Figura 3.1: Esquema de las etapas del proceso de imputación múltiple.

En la siguiente sección se analizará el método MICE que es el que será utilizado en la aplicación de esta tesis.

Técnica de Imputación Múltiple con Ecuaciones en Cadenas

Las Técnicas de Imputación Múltiple con Ecuaciones en Cadenas (MICE por su nombre en ingles Multivariate Imputation by Chained Equation) es un método de imputación múltiple donde el algoritmo 2 presenta los pasos usados para la fase de imputación de este método.

Algoritmo 2 Fase de imputación del método MICE

1: Se especifica un modelo de imputación $P(\boldsymbol{Y}_{F,j}|\boldsymbol{Y}_{O,j}, \boldsymbol{Y}_{-j}, \boldsymbol{M})$ para la variable $\boldsymbol{Y}_j$ con $j = 1, \ldots, p$.
2: Para cada j, se llena con una imputación inicial $\boldsymbol{Y}_j^0$ por extracción aleatoria de $\boldsymbol{Y}_{j,O}$.

3: **for** $i = 1, \ldots, T$ **do**
4: **for** $j = 1 \ldots, p$ **do**
5: Definimos $\boldsymbol{Y}_{-j}^i = (\boldsymbol{Y}_1^i, \ldots, \boldsymbol{Y}_{j-1}^i, \boldsymbol{Y}_{j+1}^{i-1}, \ldots, \boldsymbol{Y}_p^{i-1})$ como el actual conjunto de datos completo sin la variable $\boldsymbol{Y}_j$.
6: Sacamos $\boldsymbol{\theta}_j^i \sim P(\boldsymbol{\theta}_j^i|\boldsymbol{Y}_{O,j}, \boldsymbol{Y}_{-j}, \boldsymbol{M})$.
7: Sacamos la imputación $\boldsymbol{Y}_j^i \sim P(\boldsymbol{Y}_{F,j}|\boldsymbol{Y}_{O,j}, \boldsymbol{Y}_{-j}, \boldsymbol{M}, \boldsymbol{\theta}_j^i)$
8: **end for**
9: **end for**

El algoritmo de MICE es una Cadena de Markov de Monte Carlo, donde el espacio de estados es la colección de todos los valores imputados.

Se observa en simulaciones usando cantidades moderadas de datos faltantes que sólo 5 a 10 iteraciones son necesarias para este método [10], sin embargo, varios estudios sugieren usar un número mayor para asegurar la convergencia. La diferencia en que este método converja de manera más rápida que otros métodos de Cadenas de Markov de Monte Carlo se debe a que los datos imputados tienen una gran cantidad de aleatoriedad, la cual depende de la relación entre las variables, por lo tanto, la autocorrelación sobre el número de iteración será bajo, y esto hará una convergencia más rápida. Una alta correlación entre variables, grandes cantidades de datos faltantes y restricciones sobre los parámetros a través de diferentes variables son algunas situaciones en las que se debe tener en cuenta para el número de iteraciones. Por lo tanto, para observar la convergencia se gráfica uno o más estadísticos de interés (la media, la desviación estándar y la correlación entre algunas variables) entre iteraciones contra el número de iteración, estas deben ser libres de alguna tendencia, y la varianza dentro de una cadena debe aproximadamente ser la varianza entre cadenas. Una debilidad teórica del algoritmo MICE es que las condiciones bajo las cuales converge son desconocidos.

3.3.5. Regla de Rubin

Para la etapa de combinación, se utilizan las reglas de Rubin [11]. Sea $\boldsymbol{\theta}_k$ y $\boldsymbol{V}_k$ los estimadores y varianza respectivamente obtenidos del conjunto de datos imputado

k (para $k = 1, \ldots, M$). El estimador combinado es el promedio de los estimadores individuales:

$$\bar{\boldsymbol{\beta}} = \frac{1}{M} \sum_{i=1}^{M} \boldsymbol{\beta}_i \tag{3.4}$$

La varianza dentro de las imputaciones se estima por medio de la siguiente ecuación:

$$\bar{\boldsymbol{V}} = \frac{1}{M} \sum_{i=1}^{M} \boldsymbol{V}_i \tag{3.5}$$

La varianza entre imputaciones se aproxima por medio de:

$$\boldsymbol{B} = \frac{1}{M-1} \sum_{i=1}^{M} (\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}})(\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}})^T \tag{3.6}$$

La varianza total de $\bar{\boldsymbol{\beta}}$ se deduce de las ecuaciones 3.5 y 3.6 y es dada por la ecuación:

$$Var(\bar{\boldsymbol{\beta}}) = \bar{\boldsymbol{V}} + (1 + M^{-1})\boldsymbol{B} \tag{3.7}$$

La siguiente ecuación nos da una aproximación para calcular los intervalos de confianza,

$$\bar{\boldsymbol{\beta}} \pm \boldsymbol{t}_{\boldsymbol{\nu}} \sqrt{Var(\bar{\boldsymbol{\beta}})}^2 \tag{3.8}$$

donde $\boldsymbol{t}_{\boldsymbol{\nu}}$ es un vector que contiene los percentiles apropiados de la distribución $t-$Student central y $\boldsymbol{\nu}$ es un vector que contiene los grados de libertad que son estimados por:

$$\boldsymbol{\nu} = (M-1)\left\{1 + \frac{\bar{\boldsymbol{V}}}{(1 + \frac{1}{M})\boldsymbol{B}}\right\} \tag{3.9}$$

Capítulo 4

Aplicación a la base de datos de Cáncer de Mama

En este capitulo se expone el análisis al que fue sometida una base de datos de pacientes diagnosticadas con cáncer de mama con el objetivo de encontrar un biomarcador predictivo, es decir, identificar subpoblaciones de pacientes con la más alta probabilidad de eficacia a responder a un tratamiento contra la enfermedad por medio de la cuantificación de su severidad y haciendo una distinción de pacientes con diferentes niveles de riesgo.

En el primer apartado se da una descripción de los datos y de las variables que se van a ocupar para la definición del biomarcador. Posteriormente se aplica el método de imputación múltiple con ecuaciones encadenadas donde se ajusta el modelo de Larson & Dinse (1985) que es ocupado para definir el biomarcador predictivo. Después se evalúa la bondad de ajuste del modelo y por último se dan los resultados que se encontraron al ajustar este modelo.

4.1. Descripción de los datos

Los datos que fueron analizados corresponden a un seguimiento de 20 años de la base de datos del programa Surveillance, Epidemiology, and End Results del National Cancer Institute (SEER) de Estados Unidos, cuyos registros recolectan rutinariamente información sobre pacientes de cáncer de mama recién diagnosticadas quienes residen en Alaska, California, Connecticut, Georgia, Hawaii, Iowa, Michigan, Nuevo México, Utah y Washington (estos estados condensan el 26 % del total de la población de Estados Unidos de América). La información contenida en los archivos del SEER incluye características sociodemográficas y clinicopatológicas del tumor así como el sitio especifico de la cirugía, si se recibió radiación posoperatoria, la fecha de diagnóstico del cáncer, así como la fecha y causa de muerte.

Se tomaron registros sólo de mujeres diagnosticadas con cáncer de mama primario unilateral, es decir, que el cáncer se desarrolló en una sola mama ya sea izquierda o

derecha, y sometidas a cirugía en 1990 con fecha final de seguimiento del 31 de diciembre de 2011. Se excluyeron los casos que fueron diagnosticados por certificado de muerte o autopsia y con cirugía en otros lugares o lugares alejados de la mama y de los ganglios linfáticos. Las variables consideradas en este estudio se enlistan a continuación.

- RE: Contiene el receptor de estrógeno de la paciente tomando dos categorías, "positivo" o "al límite" ($>= 10\,\%$) y "negativo" ($< 10\,\%$).
- HS: Categoriza los hábitos de salud de la paciente con respecto al America's Health Ranking [12] en 1990, tomando como lugar el Estado en el que la paciente fue diagnosticada. Se codificó con dos categorías, las menores o iguales a 0.5 y las mayores a 0.5.
- EDAD: Indica la edad de la paciente cuando fue diagnosticada con cáncer de mama, esta variable es numérica.
- RAZA: Es la raza a la cual pertenece, tomando como categorías la raza blanca, la raza negra y otro (los indios americanos/ nativos de Alaska, Asia, Islas del Pacífico).
- ESTADO CIVIL: Se eligen las categorías de solteras (nunca casadas, separadas, divorciadas o viudas) y las casadas (incluido el derecho común).
- TIPO: Es el tipo de cáncer. Se eligen tres categorías: el carcinoma ductal, el lobulillar y otro.
- TAMAÑO: Es el tamaño del tumor tomando dos categorías, los menores a 2 cm y los mayores o iguales a 2 cm.
- GRADO: Se eligen dos categorías, una con grado I y II y la segunda con grado III y IV.
- LATERALIDAD: El lugar donde se desarrolló el cáncer de mama, las categorías son izquierda y derecha.
- EXTENSIÓN: Esta variable nos indica si el cáncer es confinado o invasivo.
- GANGLIOS LINFÁTICOS: Se toman dos categorías: si fueron o no afectados los ganglios linfáticos.
- RADIACIÓN: Esta variable indica a las mujeres que se sometieron a radioterapia. Se toman dos categorías: las mujeres que se sometieron después o antes de la cirugía y las que no tuvieron este tratamiento.
- CIRUGÍA: Esta variable indica a qué tipo de cirugía se sometieron. Para la primera categoría se trataron con cirugía conservadora de seno, en la segunda con una mastectomía.

En receptor de estrógeno se juntan las dos categorías "positivo" y "al límite" ya que en anteriores estudios se ha demostrado que tienen características similares [13]. El estado civil se categoriza de esta manera debido ya que se ha observado que pacientes nunca casadas, separadas, divorciadas o viudas tienen un mayor riesgo de morir del cáncer que las mujeres casadas [14]. Las variables que indican el tipo de tratamiento (radiación y cirugía) se incluyen en este estudio ya que aumenta significativamente la supervivencia de la paciente [15]. Asimismo, se ha observado que la edad, raza, el tipo de cáncer, el tamaño del tumor y el grado tienen un efecto significativo en la supervivencia del paciente [16]. También se ha encontrado una relación del estado de afectación de los ganglios linfáticos en la supervivencia [17]. La extensión de la enfermedad y el grado son importantes ya que dependiendo del tratamiento se puede tener una importante ventaja en la supervivencia del paciente [18]. Ha existido el interés en el hecho de que mujeres con cáncer en la mama izquierda con tratamiento de radioterapia podría dar lugar a muerte debido a problemas cardíacos, por esta razón la variable lateralidad es incluida en el estudio [19]. La variable que tiene la información de los hábitos de salud fue tomada en cuenta para observar si afectan en la supervivencia de las pacientes.

Para describir los datos se realiza la tabla de contingencia 4.1 según la cirugía y radiación. Se observa que existen datos faltantes. Para la variable ESTADO CIVIL se tiene un 2.46 %, TAMAÑO 12.23 %, GRADO 53.19 %, EXTENSIÓN 1.25 %, GANGLIOS LINFÁTICOS 7.04 % y RE 33.59 % estos serán tratados en la sección 4.2. Las proporciones de pacientes que se sometieron a cirugía conservadora de seno y que recibieron o no radioterapia posoperatoria son similares, pero los pacientes que recibieron radiación tuvieron una mejor supervivencia al final del seguimiento. Por otro lado, las que se sometieron a mastectomía y radiación posoperatoria superan a más del doble en muerte de cáncer de mama a aquellas que solamente se sometieron a mastectomía, teniendo así la tasa de supervivencia más baja del estudio.

Las mujeres que no se sometieron a radiación fueron en su mayoría de 59 años de edad o más, mientras que las mujeres con radiación posoperatoria fueron menos frecuentes en el grupo de 71+ años de edad sugiriendo que la radioterapia se administró más discriminadamente entre los pacientes de edad avanzada. En su mayoría las mujeres fueron de etnia blanca y casada. El carcinoma ductal fue el tipo de cáncer más presentado. El cáncer confinado al seno fue el que se presentó con más frecuencia. La proporciones de tamaños de tumores mayores o iguales a 2 cm fueron similares con y sin radiación después de la cirugía conservadora de seno; sin embargo, entre las mujeres que tuvieron mastectomía la proporción fue mayor para las que se sometieron a una radiación posoperativa. Los tumores de las pacientes con radioterapia posoperatoria aparecieron con más proporción a ser de mayor grado histológico y el receptor de estrógeno positivo o en el límite. La proporción en la lateralidad en los cuatro tipos de tratamientos es muy similar. La mayoría de las pacientes que se sometieron a la cirugía conservadora de seno no tenían afectación en los ganglios linfáticos axilares mientras que la afectación en los ganglios linfáticos fue más frecuente para tratamientos con radiación y mastectomía.

La figura 4.1 muestra el histograma y su curva de densidad de las razones de incidencia para los tipos de cáncer ductal, lobulillar y otros. Las gráficas muestran densidades

	Conservación de seno, (n = 5670, 34.3%)		Mastectomía, (n = 10841, 65.7%)	
Características	Sin radiación, n(%)	Con radiación, n(%)	Sin radiación, n(%)	Con radiación, n(%)
	2415 (42.6)	3255 (57.4)	9981 (92.1)	860 (7.9)
Estado vital				
Vivo	816 (33.8)	1490 (45.8)	3513 (35.2)	200 (23.3)
Muerte por cáncer de mama	430 (17.8)	536 (16.5)	2201 (22.0)	455 (52.9)
Muerte por otras causas	1169 (48.4)	1229 (37.7)	4267 (42.8)	205 (23.8)
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/etnicidad				
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)
Otro	110 (4.5)	154 (4.7)	552 (5.5)	62 (7.2)
Estado civil				
Sin pareja	1100 (45.5)	1126 (34.6)	4126 (41.3)	339 (39.4)
Casada	1197 (49.6)	2074 (63.7)	5641 (56.5)	501 (58.3)
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)
Características del tumor				
Tipo				
Carcinoma ductal	1802 (74.6)	2708 (83.2)	8099 (81.1)	662 (77.0)
Lobulillar	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 del tumor				
< 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 del tumor (RE)				
Positivo o al límite	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				
Sin afectación	1556 (64.4)	2523 (77.5)	6768 (67.8)	194 (22.6)
Con afectación	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 4.1: Tabla de contingencia según tratamiento quirúrgico y radioterapia.

bimodales con puntos de inflexión cercanas a la edad de la menopausia lo cual podría sugerir dos tipos de subpoblaciones. Se ha observado que el cáncer de mama diagnosticado a mujeres jóvenes tiende a tener un peor pronóstico que para las mujeres más grandes [20], [21], [22], [23]. Entonces, se toma la hipótesis de que la mortalidad varía con respecto a la edad para la incidencia y el tipo de muerte condicionado. En consecuencia, se ajusta un polinomio ortogonal de segundo grado para la variable edad.

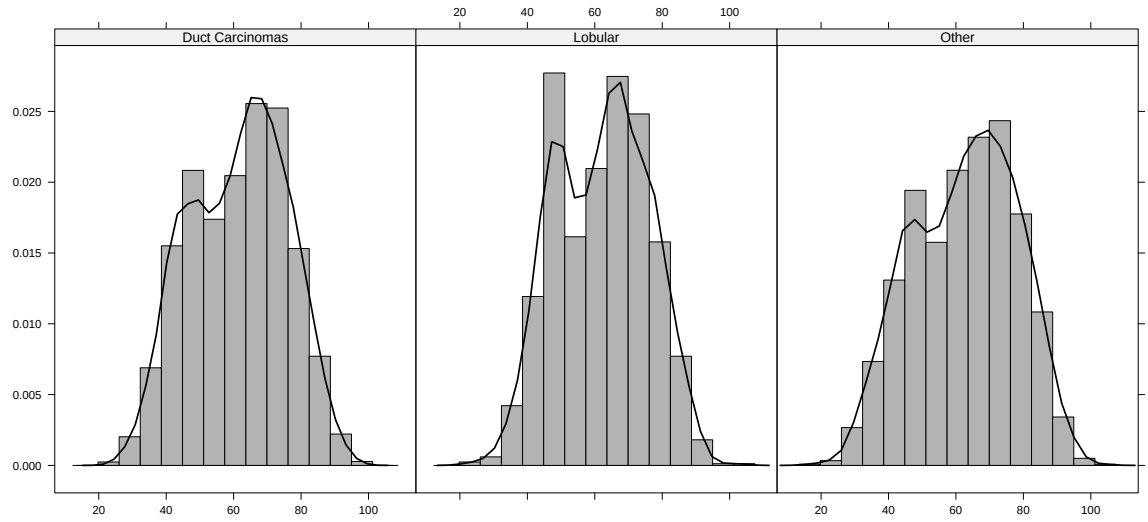


Figura 4.1: Histograma y curva de densidad de la edad de diagnostico para ductal, lobulillar y otros tipos de carcinoma.

4.2. Tratamiento de datos faltantes

Como se puede observar en la tabla 4.1 algunas variables tenían datos faltantes. Para tratarlos se recurrió al método de imputación múltiple por ecuaciones encadenadas (MICE) bajo la suposición que el mecanismo de datos faltantes es MAR; es decir, la probabilidad que un dato sea faltante solamente depende de las variables que se están utilizando en el estudio. En esta sección se presentan los pasos que se siguieron para este método (para una explicación del método se remite al lector a la sección 3.3.4).

4.2.1. Etapa de imputación

Como primer paso se eligió como modelo de imputación el modelo de regresión logística multinomial debido a que se tienen variables cualitativas con dos o más categorías. Ya habiendo definido el modelo de imputación se utilizó el análisis de datos completos con el fin de buscar el modelo más parsimonioso por medio del Criterio de Información Bayesiano con la función `step` del software estadístico `R`. Se incluyó el mayor número

de variables para minimizar el sesgo y hacer la suposición MAR más plausible [24]. A continuación se muestran los modelos más parsimoniosos encontrados por este método para las variables con datos faltantes.

- RE ∼ polinomio(EDAD, 1) + TIPO + RAZA + GRADO + TAMAÑO + LATERALIDAD + ESTADO CIVIL + HS + CIRUGÍA + LATERALIDAD:ESTADO CIVIL + ESTADO CIVIL:CIRUGÍA.

- GRADO ∼ polinomio(EDAD, 2) + SITIO + RAZA + RE + EXTENSIÓN + GANGLIOS LINFÁTICOS + TAMAÑO + LATERALIDAD + HS + CIRUGÍA + RADIACIÓN + SITIO:RE + SITIO:GANGLIOS LINFÁTICOS + EXTENSIÓN:GANGLIOS LINFÁTICOS + EXTENSIÓN:CIRUGÍA + TAMAÑO:HS + LATERALIDAD:CIRUGÍA + LATERALIDAD:RADIACIÓN + HS:CIRUGÍA

- EXTENSIÓN ∼ polinomio(EDAD, 1) + TIPO + RAZA + GRADO + NODOS LINFÁTICOS + TAMAÑO + HS + CIRUGÍA + RADIACIÓN + RAZA:NODOS LINFÁTICOS + GRADO:NODOS LINFÁTICOS + GRADO:CIRUGÍA + GRADO:RADIACIÓN + NODOS LINFÁTICOS:HS + CIRUGÍA:RADIACIÓN

- NODOS LINFÁTICOS ∼ polinomio(EDAD, 2) + TIPO + RAZA + RE + GRADO + EXTENSIÓN + TAMAÑO + HS + TRATAMIENTO + RADIACIÓN + polinomio(EDAD, 2):CIRUGÍA + TAMAÑO:SITIO + RE:TAMAÑO + RE:CIRUGÍA + GRADO:EXTENSIÓN + EXTENSIÓN:HS + RADIACIÓN:CIRUGÍA

- TAMAÑO ∼ polinomio(EDAD, 2) + SITIO + RAZA + RE + GRADO + EXTENSIÓN + ESTADO CIVIL + HS + CIRUGÍA + RADIACIÓN + NODOS LINFÁTICOS + TIPO:NODOS LINFÁTICOS + RE:NODOS LINFÁTICOS + RADIACIÓN:CIRUGÍA

- ESTADO CIVIL ∼ polinomio(EDAD, 2) + RAZA + RE + TAMAÑO + LATERALIDAD + HS + CIRUGÍA + RADIACIÓN + polinomio(EDAD, 2):RADIACIÓN + RE:LATERALIDAD + RE:CIRUGÍA

El siguiente paso fue definir el número de iteraciones e imputaciones requeridas para la convergencia del método y el ajuste, respectivamente. Para el número de iteraciones se ha observado que con 5 a 10 se alcanza la convergencia del método, sin embargo, se sugiere usar un mayor número para asegurar la convergencia [25], por lo cual se eligieron 20 iteraciones. Aunque varios estudios muestran que de 5 a 20 imputaciones es una cantidad suficiente para aplicar el método de imputación múltiple, se ha visto por medio de simulaciones que para que el error de estimación del *p-value* sea menor a 0.005 se deben de elegir al menos 100 imputaciones [26], por lo tanto, se eligieron 100 imputaciones.

Una vez teniendo los modelos más parsimoniosos, el número de imputaciones e iteraciones requeridas, se procedió a implementar el algoritmo MICE con el paquete

estadístico `mice` de `R` [27] logrando así 100 conjuntos de datos completos. Para la verificación de la convergencia de datos categóricos se utilizan tablas de frecuencia para comprobar que los datos imputados se ajustan a los datos observados [28], estas tablas se muestran en el Apéndice B, nótese que la proporción de los datos observados y los datos imputados son similares.

4.2.2. Etapa de análisis

Al realizar un test chi-cuadrado se observó que la variable GRADO tiene una gran correlación con RE con un nivel de significancia del $\alpha = 0.01$ siendo consistente con otros resultados [29], por lo que se quitó de la etapa de análisis para evitar multicolinealidad.

Debido a que una gran proporción de individuos estuvieron vivos en el fin del seguimiento, se procedió a ajustar el modelo de Larson & Dinse a cada una de las 100 imputaciones tomando como función de supervivencia base una Weibull debido a su flexibilidad en otros estudios similares de supervivencia [30]. Los estimadores fueron encontrados por el método de máxima verosimilitud usando la función `nlm` del software estadístico `R` [31] y se utilizó el hessiano para estimar la matriz de covarianza asintótica. Las variables ficticias se tomaron con nivel base igual a la primer categoría que aparece en la tabla 4.1 tomándolas igual a 0. La estrategia utilizada para encontrar el modelo más parsimonioso fue usar el Criterio de Información Bayesiano con eliminación hacia atrás en cada una de las 100 imputaciones por separado obteniendo así 100 modelos con sus respectivos conjuntos de predictores y se dejaron como predictores finales los que aparecieron en al menos la mitad de los modelos [32].

Aunque la interacción entre el polinomio de edad con cirugía y con radiación aparecieron en el componente de supervivencia condicional del mejor modelo, tales interacciones eran significativas de forma artificial, lo cual era indicativo de variables espurias, por lo tanto fueron removidas para obtener el modelo más parsimonioso.

4.2.3. Etapa de combinación

Para la etapa de combinación se utilizó la regla de Rubin dando los parámetros estimados del modelo que se eligió en la etapa de análisis que se muestran en la tabla 4.3 para los componentes de la función logística y para la condicionada y en la tabla 4.4 para la función base.

δ	β
Intercepto	Polinomio de 2° grado de edad
Polinomio de 2° grado de edad	Raza
Raza	Estado civil
Tamaño Del Tumor	Tipo
Extensión	Tamaño
Afectación Nodos Linfáticos	Receptor de estrógeno
Cirugía	Extensión
Radiación	Afectación Nodos Linfáticos
Cirugía:Radiación	Cirugía
Radiación:Extensión	Radiación
Radiación:Afectación Nodos Linfáticos	Radiación:Extensión
—	Cirugía:Radiación
—	Radiación:Extensión

Tabla 4.2: Conjunto de predictores para el modelo final.

	δ		β_1		β_2	
Intercepto	− 2.017	$(0.071)^{***}$	− − −	− − −	− − −	− − −
Polinomio Ortogonal						
1^{er}Grado	−24.856	$(2.780)^{***}$	30.235	$(2.824)^{***}$	163.368	$(2.846)^{***}$
2^{ndo}Grado	− 8.036	$(2.711)^{**}$	26.371	$(2.486)^{***}$	− 4.670	$(2.133)^{*}$
Raza (Blanca como base)						
Negra	0.288	$(0.074)^{***}$	0.228	$(0.065)^{***}$	0.279	$(0.050)^{***}$
Otro	− 0.344	$(0.100)^{***}$	0.010	(0.101)	− 0.182	$(0.064)^{**}$
Casada	− − −	− − −	− 0.070	(0.045)	− 0.176	$(0.027)^{***}$
Tipo (Ductal como base)						
Lobulillar	− − −	− − −	− 0.229	$(0.077)^{**}$	− 0.012	(0.046)
Otro	− − −	− − −	0.137	$(0.065)^{*}$	0.026	(0.037)
Tamaño $\geq$ 2cm	0.598	$(0.049)^{***}$	0.262	$(0.052)^{***}$	0.170	$(0.029)^{***}$
RE Negativo	− − −	− − −	0.604	$(0.053)^{***}$	− 0.009	(0.038)
Extensión Invasiva	1.295	$(0.125)^{***}$	1.170	$(0.111)^{***}$	0.415	$(0.111)^{***}$
Afectación Nodos Linfáticos	0.927	$(0.094)^{***}$	0.308	$(0.051)^{***}$	0.277	$(0.034)^{***}$
Mastectomía	− 0.034	(0.078)	− 0.002	(0.083)	− 0.141	$(0.035)^{***}$
Radiación	− 0.017	(0.083)	− 0.218	$(0.095)^{*}$	− 0.174	$(0.043)^{***}$
Mastectomía:Extensión Inv.	− 0.672	$(0.144)^{***}$	− 0.603	$(0.116)^{***}$	− 0.263	$(0.116)^{*}$
Mastectomía:Radiación	0.528	$(0.116)^{***}$	0.415	$(0.111)^{***}$	0.214	$(0.090)^{*}$
Radiación:Extensión Inv.	− − −	− − −	− 0.268	$(0.107)^{*}$	0.054	(0.125)
Radiación:Afectación Nodos	0.302	$(0.107)^{**}$	− − −	− − −	− − −	− − −
$*$ *p-value* < 0.05	$**$ *p-value* < 0.01		$***$ *p-value* < 0.001			

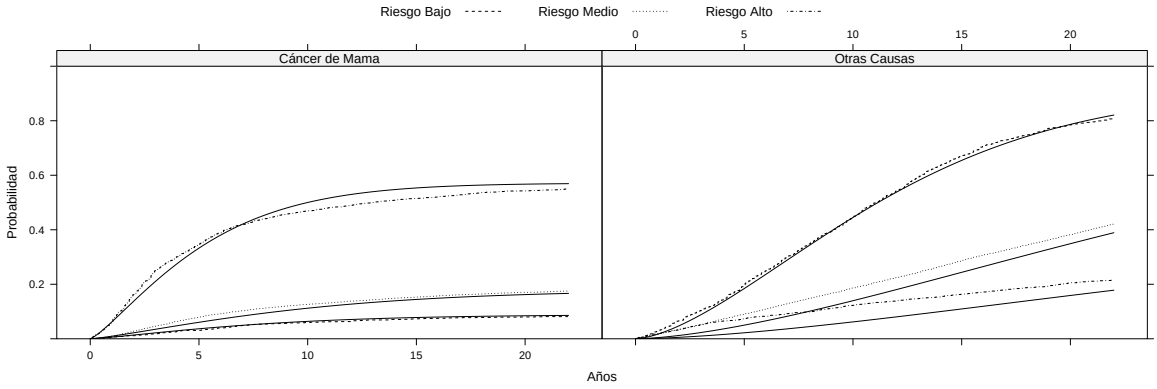
Tabla 4.3: Estimadores y errores estándar resultantes de la etapa de combinación y su *p-value* con el estadístico de Wald

$\log(\lambda_1)$		$\log(\lambda_2)$		$\log(\gamma_1)$		$\log(\gamma_2)$	
4.941	$(0.076)^{***}$	5.845	$(0.034)^{***}$	0.226	$(0.015)^{***}$	0.436	$(0.011)^{***}$
$*$ *p-value* < 0.05		$**$ *p-value* < 0.01			$***$ *p-value* < 0.001		

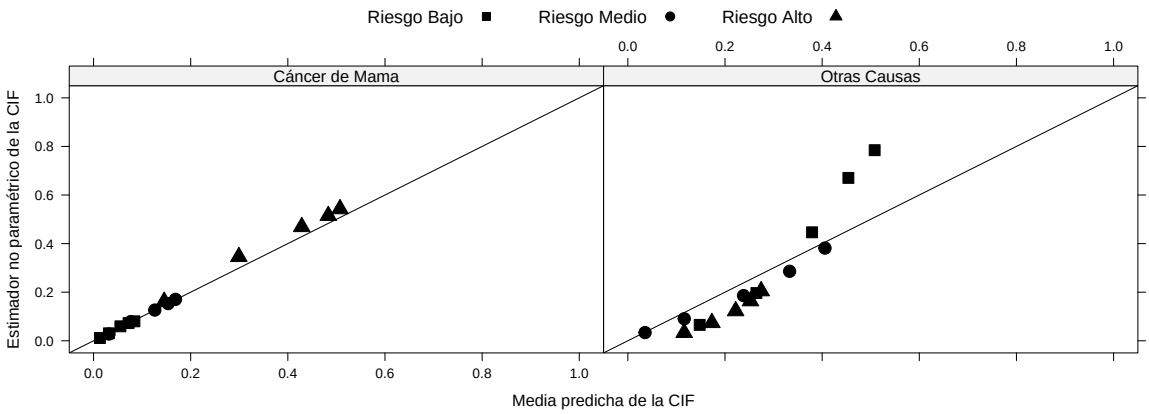
Tabla 4.4: Estimadores y errores estándar resultantes de la etapa de combinación de las funciones de supervivencia base y su *p-value* con el estadístico de Wald

4.3. Evaluación

Con el modelo propuesto se puede predecir la mortalidad por cáncer de mama y por otras causas, además es capaz de asignar diferentes grupos de riesgos a los pacientes que eventualmente podrían experimentar el evento de interés, esto se realiza mediante la puntuación de riesgo $R = \delta + \boldsymbol{\delta}^T \mathbf{z}$ donde los parámetros δ y $\boldsymbol{\delta}$ son los mostrados en la tabla 4.3. Se tomó el 10 % más bajo como el grupo de riesgo bajo, el siguiente 75 % como el grupo de riesgo medio y el 15 % restante como el grupo de riesgo alto. Dados estos grupos, se utilizaron dos medidas importantes para la evaluación del ajuste del modelo que son la validación y calibración. Para la validación se calcula la función de incidencia acumulativa de nuestro modelo y se grafica con la función de incidencia acumulativa no paramétrica del estimador de Nelson-Aalen, dando así las siguientes gráficas:



Para la calibración se utilizó una gráfica propuesta por [33] que utiliza la estimación no paramétrica de la función de incidencia acumulativa de la causa j contra la media de la función de incidencia acumulativa de los sujetos del grupo de riesgo de la causa j con $j = 1, 2$. Para que un modelo sea bueno, la gráfica debe de pertenecer a una línea recta identidad, se grafican los puntos de los años 2, 5, 10, 15 y 20.



4.4. Resultados

Debido a que pacientes con peores características en el diagnostico fueron más probables a someterse a mastectomía o radioterapia posoperativa o a ambos, cualquier comparación entre tratamientos podrían confundir por la diferencia de la severidad de cáncer de mama entre pacientes y esto lo hace altamente susceptible a sesgo, por lo tanto, las variables CIRUGÍA y RADIACIÓN fueron tomadas como variables de confusión. El método utilizado para tratar a las variables de confusión es interpretar los efectos significativos de cirugía, radiación y variables con interacción de cirugía y radiación como efectos causales, y a los demás factores en el análisis de regresión como efectos de riesgo. La tabla 4.3 muestra el conjunto final de predictores, la cual incluye la interacción entre el tipo de cirugía y la radiación en el modelo logístico y condicional, también la extensión de la enfermedad y las interacciones con cirugía y radiación, mientras que la interacción de afectación en los nodos linfáticos fue significativa sólo en el modelo logístico. Lo anterior es consistente con el hecho que la elección de la cirugía y de la radioterapia es basada en la información del diagnóstico siendo tratada con radioterapia después de la cirugía cuando hay afectación en los nodos linfáticos. De lo anterior, cirugía, radiación, extensión y afectación en los nodos linfáticos se pueden considerar como efectos causales.

Lateralidad y hábitos de salud no tuvieron efectos significativos en ninguno de los componentes del modelo de mezclas, lo cual sugiere que ni los hábitos de salud ni las condiciones en que se vive cambian el riesgo de muerte de una causa específica y que no hay evidencia que la lateralidad está asociada con un mayor riesgo de morir de cáncer de mama. Algo interesante es el hecho de que la interacción entre lateralidad y radiación no fue importante, lo cual concuerda con investigaciones actuales que la radioterapia en el lado izquierdo del pecho como se hacía en los años 90's no incrementa el riesgo de morir de alguna causa cuando se compara a los tumores del lado derecho [19], [34].

La tabla 4.3 muestra los estimadores del modelo de mezclas más parsimonioso para los componentes de la función logística y para la condicionada. A continuación se presentan los resultados más significativos con un intervalo de confianza del 95 %.

Los coeficientes que corresponden a la raza muestran evidencia de diferencias entre los tres grupos raciales consideradas en este estudio en la mortalidad de cáncer de mama. El porcentaje de mujeres de raza negra que morirán de cáncer de mama es mayor con un $15\,\% - 54\,\%$ que las mujeres de raza blanca, mientras que el de las mujeres de otras etnicidades es menor con $14\,\% - 42\,\%$ que las mujeres de raza blanca. Cuando se compara las razones de fuerza de mortalidad, las mujeres de raza negra tienen mayor proporción de morir de cáncer de mama o de otras causas que las mujeres de raza blanca, para el riesgo condicionado del cáncer de mama es mayor con $10\,\% - 43\,\%$ y de otras causas $20\,\% - 46\,\%$, y para las mujeres de otras etnicidades no parece haber diferencia en la condicional de cáncer de mama, pero para otras causas es un $5\,\% - 26\,\%$ menor que las mujeres de raza blanca.

Estado civil fue significativa en la razón de fuerza de mortalidad de morir de otras causas con un coeficiente negativo, dando a entender que el apoyo social tiene un impacto positivo en la condicional de supervivencia después de que se le diagnóstico de cáncer de mama, pero no tiene un efecto significativo en la condicional de morir de cáncer de mama o en la causa eventual de morir, lo cual es un cambio de perspectiva con anteriores estudios realizados, los cuales afirman que las mujeres solteras tienen un mayor riesgo de morir por el cáncer [14].

Se puede observar que la mortalidad condicionada al cáncer de mama para el carcinoma lobulillar tuvo un porcentaje de $8\,\% - 32\,\%$ menor asociado con el riesgo de mortalidad condicionado al cáncer de mama comparado con el carcinoma ductal, mientras que otros subtipos tuvieron un porcentaje del $1\,\% - 30\,\%$ mayor comparado con el carcinoma ductal; finalmente, la mortalidad condicionada a otras causas parece no tener efecto.

El tamaño del tumor tuvo efectos muy significativos en los tres componentes de la mezcla. Para tumores $\geq$ 2cm se obtuvo un $65\,\% - 100\,\%$ en la mortalidad de cáncer de mama, para el riesgo condicional de muerte por cáncer de mama un $17\,\% - 44\,\%$ y para muerte condicionada a otras causas $12\,\% - 25\,\%$ mayor a tumores $<$ 2cm, lo que podría reflejar los efectos secundarios endógenos que encarnan tumores grandes, ya que se puede especular que los tratamientos adyuvantes más agresivos son dados a pacientes con tumores más grandes en los años 90's.

La variable de RE sólo tuvo efectos significativos en la condicional de cáncer de mama, indicando que cuando es negativo entonces se es asociado a un $65\,\% - 103\,\%$ más grande de riesgo de mortalidad comparado con el positivo o al límite.

El polinomio ortogonal de edad fue significativo tanto en la componente logística como en las funciones condicionadas. La figura 4.2 muestra las curvas del polinomio de edad de diagnóstico con una banda de confianza del 95 % en los tres componentes de la mezcla.

La primera gráfica corresponde al efecto en el componente logístico. Como se puede observar hay una amplia anchura en las bandas de confianza para pacientes diagnosticadas antes de la perimenopausia y gradualmente se vuelve más estrecha para las edades dentro de este rango, en este punto se podría superponer una función constante entre las bandas de confianza, revelando que este grupo de pacientes comparten un riesgo

similar de morir eventualmente de cáncer de mama. Después el riesgo disminuye para pacientes con edad después de la menopausia y las bandas de confianza continúan estrechas hasta la edad de los 80's y se expande para pacientes de mayor edad mostrando una trayectoria decreciente, lo que indica que la probabilidad de morir de cáncer de mama decrementó como la edad incrementa en este grupo de mujeres, lo cual confirma el hecho que mujeres jóvenes son más probables de morir del cáncer de mama [20].

La gráfica de en medio muestra que un paciente muera eventualmente de cáncer de mama, la razón de fuerza de mortalidad disminuye como la edad incrementa para pacientes diagnosticados antes de la premenopausia, y entonces incrementa para pacientes diagnosticados después de la menopausia.

Por último, la gráfica de la derecha indica que un paciente muera eventualmente por otras causas, la razón de fuerza de mortalidad incrementa con la edad, como es de esperarse, lo cual se observa por la anchura de las bandas de confianza.

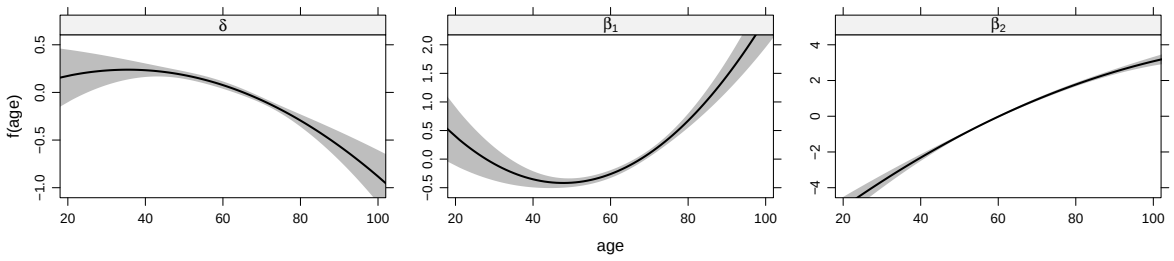


Figura 4.2: Polinomio ortogonal de la edad de diagnóstico con una banda de confianza del 95 % en el componente logístico (δ), en el componente condicional de morir de cáncer de mama (β_1) y el componente condicional de morir de otras causas (β_2) para el modelo de mezclas más parsimonioso.

Capítulo 5

Conclusiones

En el presente trabajo se implemento un modelo de mezclas para riesgos competitivos para el análisis de regresión de la mortalidad por causas específicas a largo plazo en presencia de elementos sociodemográficos y clínicopatológicos dentro de una cohorte de base poblacional a gran escala de mujeres diagnosticadas con cáncer de mama que se sometieron a cirugía. Se clasificaron dos tipos de muerte, uno por otras causas y otro por el cáncer de mama, bajo esta formulación existieron tres coeficientes, uno que describe cómo el predictor afecta la mortalidad por causas específicas a largo plazo y dos que describen cómo afecta a cada causa de muerte, lo cual permite diversas interpretaciones que se presentan para cada factor. Bajo esta metodología se logró la exploración y definición de un biomarcador predictivo en términos de las características importantes de un paciente, que resultó ser una herramienta precisa tanto para la cuantificación de la severidad de la enfermedad y la discriminación de pacientes con diferentes niveles de riesgo de la eventual muerte por cáncer de mama. Con este modelo fue posible confirmar anteriores estudios epidemiológicos así como entender aparentes contradicciones en estudios anteriores. Las conclusiones con respecto a los objetivos planteados son:

- El modelo de mezclas de Larson & Dinse (1985) para el análisis de riesgos competitivos permitió incluir variables explicativas tanto sociodemográficas como clinopatológicas.

- Con el método MICE se logró incluir datos de pacientes cuyos registros presentaban al menos una variable no observada por medio de la imputación (sustitución).

- La ventaja del modelo propuesto en este trabajo es que las inferencias basadas en verosimilitud son muy parecidas a los modelos generalizados, por lo tanto, fue posible aplicarlos a los conjuntos de datos completos obtenidos al aplicar el método de imputación múltiple con ecuaciones encadenadas.

- El marcador predictivo fue obtenido por medio de la puntuación de riesgo $R = \delta + \boldsymbol{\delta}^T \mathbf{z}$ discriminando pacientes con diferentes niveles de riesgo (bajo, medio y

alto) de morir de cáncer de mama y cuantificar la severidad de la enfermedad en términos de características de los pacientes.

- Se utilizaron dos medidas importantes para la evaluación del ajuste, la validación y la calibración, las cuales mostraron que el modelo utilizado en esta aplicación es un modelo adecuado.

Apéndice A

Función de log verosimilitud del modelo de Larson & Dinse (1985) con condicionales Weibull en R

```
weib.mix <- function(parametros, tipo, Ti, matriz.p, matriz.S)
{
  n.var.p <- dim(matriz.p)[2]
  n.var.S <- dim(matriz.S)[2]
  J <- max(tipo)
  nt <- length(tipo)
  C.ij <- matrix(1:(nt * J), ncol = J) * 0
  for(i in 1:J)
  {
    C.ij[, i][tipo == i] <- 1
  }
  uno <- rep(1, J)
  ci <- C.ij %*% uno
  Tij <- matrix(1:(nt * J), ncol = J) * 0 + 1
  T.ij <- Tij * Ti
  deltas <- parametros[1:((J-1)*n.var.p)]

  if(n.var.S == 1) {
    betas <- rep(0, J)
    log.lambdas <- parametros[((J-1)*n.var.p+1):
                              ((J-1)*n.var.p+J)]
    log.shapes <- parametros[((J-1)*n.var.p+J+1):
                             ((J-1)*n.var.p+J+J)]
    Vij <- matriz.S[,1]
  } else if(n.var.S > 1) {
```

```
  betas <- parametros[((J-1)*n.var.p+1):
                        ((J-1)*n.var.p+(n.var.S-1)*J)]
  log.lambdas <- parametros[(((J-1)*n.var.p+(n.var.S-1)*J)+1):
                             (((J-1)*n.var.p+(n.var.S-1)*J)+J)]
  log.shapes <- parametros[(((J-1)*n.var.p+(n.var.S-1)*J)+J+1):
                           (((J-1)*n.var.p+(n.var.S-1)*J)+J + J)]
  Zij <- matriz.S[,1]#Valores del vector interseccion
  Vij <- matriz.S[,-1]#Valores del ultimo vector de parametros
}

dj.u <- matrix(1:(J * nt), ncol = J) * 0
Uij <- matriz.p

if(n.var.p==1) {
  for(i in 1:(J - 1))
  {
    dj.u[, i] <- c(Uij * deltas[i])
  }
} else if(n.var.p > 1) {
  for(i in 1:(J - 1))
  {
    dj.u[, i] <- c(Uij %*% deltas[((i-1)*n.var.p+1):(i*n.var.p)])
  }
}

bj.v <- matrix(1:(J * nt), ncol = J)
lambdas <- matrix(1:(J * nt), ncol = J)
shape.ij <- matrix(rep(1,(nt * J)), ncol = J)

if(n.var.S==1) {
  for(i in 1:J)
  {
    bj.v[, i] <- c(Vij * betas[i])
    lambdas[, i] <- exp(c(matriz.S * log.lambdas[i]))
  }
} else if(n.var.S == 2) {
  for(i in 1:J)
  {
    bj.v[, i] <- c(Vij * betas[((i-1)*(n.var.S-1) + 1):
                               (i*(n.var.S-1))])
    lambdas[, i] <- exp(c(Zij* log.lambdas[i]))
    shape.ij[, i] <- exp(c(rep(log.shapes[i], nt)))
  }
```

```
  } else if(n.var.S > 2) {
    for(i in 1:J)
    {
      bj.v[, i] <- c(Vij %*% betas[((i-1)*(n.var.S-1) + 1):
                                   (i*(n.var.S-1))])
      lambdas[, i] <- exp(c(Zij* log.lambdas[i]))
      shape.ij[, i] <- exp(c(rep(log.shapes[i], nt)))
    }
  }

  p.ij <- matrix(1:(J * nt), ncol = J) * 0
  f0.ij <- matrix(1:(J * nt), ncol = J) * 0
  S0.ij <- matrix(1:(J * nt), ncol = J) * 0
  exp.dj.u <- exp(dj.u)
  sum.exps.i <- exp.dj.u %*% uno

  for(i in 1:J)
  {
    p.ij[, i] <- exp.dj.u[, i]/(sum.exps.i)
    f0.ij[, i]  <- dweibull(T.ij[, i], shape=shape.ij[, i],
                            scale = lambdas[, i])
    S0.ij[, i]  <- 1 - pweibull(T.ij[, i], shape=shape.ij[, i],
                                scale = lambdas[, i])
  }

  l.ij <- exp(bj.v)
  h0.ij <- f0.ij/S0.ij
  S.ij <- (S0.ij)^{l.ij}
  f.ij <- h0.ij*l.ij*S.ij
  A <- sum(C.ij * (log(p.ij) + log(f.ij)))
  p.S.ij <- p.ij * S.ij
  sumJ.p.S.ij <- p.S.ij %*% uno
  B <- sum((1 - ci) * log(sumJ.p.S.ij))
  log.lik <- A + B
  -1*log.lik
}
```

Apéndice B

Tablas de frecuencias de los datos imputados

Datos Observados			
	Frecuencias	Porcentaje	Porcentaje Acumulado
Estado civil			
Sin pareja	6691	41.55	41.55
Casada	9413	58.45	100
Total	16104	100	
Tamaño del Tumor			
$<$ 2 cm	8251	56.94	56.94
$\geq$ 2 cm	6240	43.06	100
Total	14491	100	
Grado			
I & II	4033	52.19	52.19
III & IV	3695	47.81	100
Total	7728	100	
Marcador del tumor (RE)			
Positivo o al límite	8437	76.94	76.94
Negativo	2528	23.05	100
Total	10965	100	
Extensión			
Confinado	14847	91.06	91.06
Invasivo	1457	8.94	100
Total	16304	100	
Ganglios linfáticos			
Sin afectación	11041	71.94	71.94
Con afectación	4306	28.06	100
Total	15347	100	

Tabla B.1: Tabla de frecuencia de los datos observados.

Imputación 10			
	Frecuencias	Porcentaje	Porcentaje Acumulado
Estado civil			
Sin pareja	6886	41.70	41.70
Casada	9625	58.30	100
Total	16511	100	
Tamaño del Tumor			
< 2 cm	9477	57.40	57.40
≥ 2 cm	7034	42.60	100
Total	16511	100	
Grado			
I & II	8908	53.95	53.95
III & IV	7603	46.05	100
Total	16511	100	
Marcador del tumor (RE)			
Positivo o al límite	12811	77.59	77.59
Negativo	3700	22.49	100
Total	16511	100	
Extensión			
Confinado	15027	91.01	91.01
Invasivo	1484	8.98	100
Total	16511	100	
Ganglios linfáticos			
Sin afectación	11873	71.91	71.91
Con afectación	4638	28.09	100
Total	16511	100	

Tabla B.2: Tabla de frecuencia de la imputación 10.

Imputación 20			
	Frecuencias	Porcentaje	Porcentaje Acumulado
Estado civil			
Sin pareja	6883	41.69	41.69
Casada	9628	58.31	100
Total	16511	100	
Tamaño del Tumor			
< 2 cm	9465	57.32	57.32
≥ 2 cm	7046	42.68	100
Total	16511	100	
Grado			
I & II	8874	53.75	53.75
III & IV	7637	46.25	100
Total	16511	100	
Marcador del tumor (RE)			
Positivo o al límite	12859	77.88	77.88
Negativo	3652	22.12	100
Total	16511	100	
Extensión			
Confinado	15022	90.98	90.98
Invasivo	1489	9.02	100
Total	16511	100	
Ganglios linfáticos			
Sin afectación	11855	71.80	71.80
Con afectación	4656	28.20	100
Total	16511	100	

Tabla B.3: Tabla de frecuencia de la imputación 20.

Imputación 30			
	Frecuencias	Porcentaje	Porcentaje Acumulado
Estado civil			
Sin pareja	6880	41.67	41.67
Casada	9631	58.33	100
Total	16511	100	
Tamaño del Tumor			
< 2 cm	9392	56.88	56.88
≥ 2 cm	7119	43.12	100
Total	16511	100	
Grado			
I & II	9189	55.65	55.65
III & IV	7322	44.35	100
Total	16511	100	
Marcador del tumor (RE)			
Positivo o al límite	12849	77.82	77.82
Negativo	3662	22.18	100
Total	16511	100	
Extensión			
Confinado	15029	91.02	91.02
Invasivo	1482	8.98	100
Total	16511	100	
Ganglios linfáticos			
Sin afectación	11858	71.82	71.82
Con afectación	4653	28.18	100
Total	16511	100	

Tabla B.4: Tabla de frecuencia de la imputación 30.

Imputación 40			
	Frecuencias	Porcentaje	Porcentaje Acumulado
Estado civil			
Sin pareja	6847	41.47	41.47
Casada	9664	58.53	100
Total	16511	100	
Tamaño del Tumor			
$<$ 2 cm	9419	57.05	57.05
$\geq$ 2 cm	7092	42.95	100
Total	16511	100	
Grado			
I & II	8963	54.28	54.28
III & IV	7548	45.72	100
Total	16511	100	
Marcador del tumor (RE)			
Positivo o al límite	12822	77.66	77.66
Negativo	3689	22.34	100
Total	16511	100	
Extensión			
Confinado	15025	91.00	91.00
Invasivo	1486	9.00	100
Total	16511	100	
Ganglios linfáticos			
Sin afectación	11856	71.81	71.81
Con afectación	4655	28.19	100
Total	16511	100	

Tabla B.5: Tabla de frecuencia de la imputación 40.

Imputación 50			
	Frecuencias	Porcentaje	Porcentaje Acumulado
Estado civil			
Sin pareja	6865	41.58	41.58
Casada	9646	58.42	100
Total	16511	100	
Tamaño del Tumor			
< 2 cm	9456	57.27	57.27
≥ 2 cm	7055	42.93	100
Total	16511	100	
Grado			
I & II	8967	54.31	54.31
III & IV	7544	45.69	100
Total	16511	100	
Marcador del tumor (RE)			
Positivo o al límite	12842	77.78	77.78
Negativo	3669	22.22	100
Total	16511	100	
Extensión			
Confinado	15031	91.04	91.04
Invasivo	1480	8.96	100
Total	16511	100	
Ganglios linfáticos			
Sin afectación	11878	71.94	71.94
Con afectación	4633	28.06	100
Total	16511	100	

Tabla B.6: Tabla de frecuencia de la imputación 50.

Imputación 60			
	Frecuencias	Porcentaje	Porcentaje Acumulado
Estado civil			
Sin pareja	6864	41.57	41.57
Casada	9647	58.43	100
Total	16511	100	
Tamaño del Tumor			
$<$ 2 cm	9434	57.14	57.14
$\geq$ 2 cm	7077	42.86	100
Total	16511	100	
Grado			
I & II	8974	54.35	54.35
III & IV	7537	45.65	100
Total	16511	100	
Marcador del tumor (RE)			
Positivo o al límite	12845	77.80	77.80
Negativo	3666	22.20	100
Total	16511	100	
Extensión			
Confinado	15035	91.06	91.06
Invasivo	1476	8.94	100
Total	16511	100	
Ganglios linfáticos			
Sin afectación	11878	71.94	71.94
Con afectación	4633	28.06	100
Total	16511	100	

Tabla B.7: Tabla de frecuencia de la imputación 60.

Imputación 70			
	Frecuencias	Porcentaje	Porcentaje Acumulado
Estado civil			
Sin pareja	6871	41.61	41.61
Casada	9640	58.39	100
Total	16511	100	
Tamaño del Tumor			
< 2 cm	9467	57.34	57.34
≥ 2 cm	7044	42.66	100
Total	16511	100	
Grado			
I & II	9096	55.09	55.09
III & IV	7415	44.91	100
Total	16511	100	
Marcador del tumor (RE)			
Positivo o al límite	12780	77.40	77.40
Negativo	3731	22.60	100
Total	16511	100	
Extensión			
Confinado	15020	90.97	90.97
Invasivo	1491	9.03	100
Total	16511	100	
Ganglios linfáticos			
Sin afectación	11870	71.89	71.89
Con afectación	4641	28.11	100
Total	16511	100	

Tabla B.8: Tabla de frecuencia de la imputación 70.

Imputación 80			
	Frecuencias	Porcentaje	Porcentaje Acumulado
Estado civil			
Sin pareja	6870	41.61	41.61
Casada	9641	58.39	100
Total	16511	100	
Tamaño del Tumor			
$<$ 2 cm	9459	57.29	57.29
$\geq$ 2 cm	7052	42.71	100
Total	16511	100	
Grado			
I & II	8966	54.30	54.30
III & IV	7445	45.70	100
Total	16511	100	
Marcador del tumor (RE)			
Positivo o al límite	12868	77.94	77.94
Negativo	3643	22.06	100
Total	16511	100	
Extensión			
Confinado	15023	90.99	90.99
Invasivo	1488	9.01	100
Total	16511	100	
Ganglios linfáticos			
Sin afectación	11885	71.98	71.98
Con afectación	4626	28.02	100
Total	16511	100	

Tabla B.9: Tabla de frecuencia de la imputación 80.

Imputación 90			
	Frecuencias	Porcentaje	Porcentaje Acumulado
Estado civil			
Sin pareja	6877	41.65	41.65
Casada	9634	58.35	100
Total	16511	100	
Tamaño del Tumor			
$<$ 2 cm	9457	57.28	57.28
$\geq$ 2 cm	7054	42.72	100
Total	16511	100	
Grado			
I & II	8954	54.23	54.23
III & IV	7557	45.77	100
Total	16511	100	
Marcador del tumor (RE)			
Positivo o al límite	12789	77.46	77.46
Negativo	3722	22.54	100
Total	16511	100	
Extensión			
Confinado	15033	91.05	91.05
Invasivo	1478	8.95	100
Total	16511	100	
Ganglios linfáticos			
Sin afectación	11860	71.83	71.83
Con afectación	4651	28.17	100
Total	16511	100	

Tabla B.10: Tabla de frecuencia de la imputación 90.

Imputación 100			
	Frecuencias	Porcentaje	Porcentaje Acumulado
Estado civil			
Sin pareja	6870	41.61	41.61
Casada	9641	58.39	100
Total	16511	100	
Tamaño del Tumor			
< 2 cm	9471	57.36	57.36
≥ 2 cm	7040	42.64	100
Total	16511	100	
Grado			
I & II	9023	54.65	54.65
III & IV	7488	44.35	100
Total	16511	100	
Marcador del tumor (RE)			
Positivo o al límite	12828	77.69	77.69
Negativo	3683	22.30	100
Total	16511	100	
Extensión			
Confinado	15024	90.99	90.99
Invasivo	1487	9.01	100
Total	16511	100	
Ganglios linfáticos			
Sin afectación	11868	71.88	71.88
Con afectación	4643	28.12	100
Total	16511	100	

Tabla B.11: Tabla de frecuencia de la imputación 100.

Bibliografía

[1] J Ferlay, I Soerjomataram, M Ervik, R Dikshit, S Eser, C Mathers, M Rebelo, DM Parkin, D Forman, and F Bray. Globocan 2012 v1. 0. *Cancer Incidence and Mortality Worldwide: IARC CancerBase [Internet]*, (11), 2013. Lyon, France: International Agency for Research on Cancer; 2013. Available from: http://globocan.iarc.fr, accessed on 07/02/2016.

[2] Sadanori Konishi and Genshiro Kitagawa. *Information criteria and statistical modeling*. Springer Science & Business Media, 2008.

[3] Anastasios Tsiatis. A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1):20–22, 1975.

[4] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

[5] Martin G Larson and Gregg E Dinse. A mixture model for the regression analysis of competing risks data. *Applied statistics*, pages 201–211, 1985.

[6] Stephan W Lagakos, Charles J Sommer, and Marvin Zelen. Semi-markov models for partially censored data. *Biometrika*, 65(2):311–317, 1978.

[7] SW Lagakos. General right censoring and its impact on the analysis of survival data. *Biometrics*, pages 139–156, 1979.

[8] Joseph L Schafer and John W Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

[9] Roderick JA Little. Regression with missing x's: a review. *Journal of the American Statistical Association*, 87(420):1227–1237, 1992.

[10] Stef Van Buuren, Jaap PL Brand, CGM Groothuis-Oudshoorn, and Donald B Rubin. Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, 76(12):1049–1064, 2006.

[11] Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 1987.

[12] Paul E Peppard, David A Kindig, Elizabeth Dranger, Amanda Jovaag, and Patrick L Remington. Ranking community health status to stimulate discussion of local public health issues: the wisconsin county health rankings. *American Journal of Public Health*, 98(2):209–212, 2008.

[13] Takayuki Iwamoto, Daniel Booser, Vicente Valero, James L Murray, Kimberly Koenig, Francisco J Esteva, Naoto T Ueno, Jie Zhang, Weiwei Shi, Yuan Qi, et al. Estrogen receptor (er) mrna and er-related gene expression in breast cancers that are 1 % to 10 % er-positive by immunohistochemistry. *Journal of Clinical Oncology*, 30(7):729–734, 2012.

[14] Ayal A Aizer, Ming-Hui Chen, Ellen P McCarthy, Mallika L Mendu, Sophia Koo, Tyler J Wilhite, Powell L Graham, Toni K Choueiri, Karen E Hoffman, Neil E Martin, et al. Marital status and survival in patients with cancer. *Journal of Clinical Oncology*, 31(31):3869–3876, 2013.

[15] Katherine A Vallis and Ian F Tannock. Postoperative radiotherapy for breast cancer: growing evidence for an impact on survival. *Journal of the National Cancer Institute*, 96(2):88–89, 2004.

[16] Jarrett Rosenberg, Yen Lin Chia, and Sylvia Plevritis. 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(1):47–54, 2005.

[17] Christine L Carter, Carol Allen, and Donald E Henson. Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer*, 63(1):181–187, 1989.

[18] Bevan Hong Ly, Georges Vlastos, Elisabetta Rapiti, Vincent Vinh-Hung, and Nam Phong Nguyen. Local-regional radiotherapy and surgery is associated with a significant survival advantage in metastatic breast cancer patients. *Tumori*, 96(6):947e54, 2010.

[19] Charles E Rutter, Anees B Chagpar, and Suzanne B Evans. Breast cancer laterality does not influence survival in a large modern cohort: implications for radiation-related cardiac mortality. *International Journal of Radiation Oncology* Biology* Physics*, 90(2):329–334, 2014.

[20] Carey K Anders, David S Hsu, Gloria Broadwater, Chaitanya R Acharya, John A Foekens, Yi Zhang, Yixin Wang, P Kelly Marcom, Jeffrey R Marks, Phillip G Febbo, et al. Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression. *Journal of Clinical Oncology*, 26(20):3324–3330, 2008.

[21] Chunyuan Fei, Lisa A DeRoo, Dale P Sandler, and Clarice R Weinberg. Menopausal symptoms and the risk of young-onset breast cancer. *European Journal of Cancer*, 49(4):798–804, 2013.

[22] Hatem A Azim Jr and Ann H Partridge. Biology of breast cancer in young women. 2014.

[23] Shitalmala Thangjam, Rajesh Singh Laishram, and Kaushik Debnath. Breast carcinoma in young females below the age of 40 years: A histopathological perspective. *South Asian journal of cancer*, 3(2):97–100, 2014.

[24] Joseph L Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997.

[25] Jaap Brand. *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*. 1999.

[26] James Carpenter and Michael Kenward. *Multiple imputation and its application*. John Wiley & Sons, 2012.

[27] Stef Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45(3), 2011.

[28] Donia Smaali Bouhlila and Fethi Sellaouti. Multiple imputation using chained equations for missing data in timss: a case study. *Large-scale Assessments in Education*, 1(1):1–33, 2013.

[29] Edwin R Fisher, C Kent Osborne, William L McGuire, Carol Redmond, William A Knight III, Bernard Fisher, George Bannayan, Arnold Walder, Ernest J Gregory, Avram Jacobsen, et al. Correlation of primary breast cancer histopathology and estrogen receptor content. *Breast cancer research and treatment*, 1(1):37–41, 1981.

[30] Kevin J Carroll. On the use and utility of the weibull model in the analysis of survival data. *Controlled clinical trials*, 24(6):682–701, 2003.

[31] R Core Team. R: A language and environment for statistical computing [internet]. vienna, austria: R foundation for statistical computing; 2013. *Document freely available on the internet at: http://www. r-project. org*, 2015.

[32] Angela M Wood, Ian R White, and Patrick Royston. How should variable selection be performed with multiply imputed data? *Statistics in medicine*, 27(17):3227–3246, 2008.

[33] Michael W Kattan, Glenn Heller, and Murray F Brennan. A competing-risks nomogram for sarcoma-specific death following local recurrence. *Statistics in medicine*, 22(22):3515–3525, 2003.

[34] Jing Bao, Ke-Da Yu, Yi-Zhou Jiang, Zhi-Ming Shao, and Gen-Hong Di. The effect of laterality and primary tumor site on cancer-specific mortality in breast cancer: a seer population-based study. *PloS one*, 9(4):e94815, 2014.