UNIVERSIDAD AUTÓNOMA METROPOLITANA-IZTAPALAPA
DIVISIÓN DE CIENCIAS BÁSICAS E INGENIERÍAS

IMPUTACIÓN MÚLTIPLE PARA RIESGOS COMPETITIVOS: ANÁLISIS DE SUPERVIVENCIA PARA PACIENTES CON CÁNCER DE PRÓSTATA

TESIS QUE PRESENTA:
MARCO ANTONIO BARRAGAN MARTÍNEZ
PARA OBTENER EL GRADO DE
MAESTRO EN CIENCIAS
MATEMÁTICAS APLICADAS E INDUSTRIALES

ASESOR: DR. GABRIEL ESCARELA PÉREZ

JURADO CALIFICADOR:

PRESIDENTE:	DR. GABRIEL ESCARELA PÉREZ	UAM-I
SECRETARIO:	DRA. BLANCA ROSA PÉREZ SALVADOR	UAM-I
VOCAL:	DR. GABRIEL A. RODRÍGUEZ YAM	UACh

MÉXICO, D.F. NOVIEMBRE 2013

UNIVERSIDAD AUTÓNOMA METROPOLITANA

UNIDAD - IZTAPALAPA

TESIS:

IMPUTACIÓN MÚLTIPLE PARA RIESGOS COMPETITIVOS: ANÁLISIS DE SUPERVIVENCIA PARA PACIENTES CON CÁNCER DE PRÓSTATA

PRESENTADA POR:

MARCO ANTONIO BARRAGAN MARTÍNEZ

PARA OBTENER EL GRADO DE:
MAESTRO EN CIENCIAS
(MATEMÁTICAS APLICADAS E INDUSTRIALES)

ASESOR:
Dr. GABRIEL ESCARELA PÉREZ

NOVIEMBRE 2013

"Las matemáticas son el alfabeto con el cual Dios ha escrito el Universo."

— Galileo Galilei (1564 - 1642)

Agradecimientos

A mis padres:
ELVIRA MARTÍNEZ PAEZ y TAURINO BARRAGAN FERNÁNDEZ
Gracias por su apoyo en todo momento, sin ustedes no estaría donde estoy, son los mejores padres del mundo.
A mis hermanos favoritos:
JOSSELYN Y LUIS
A ti
ALEJANDRA
Por las palabras de apoyo que siempre tenías para mí en los momentos difíciles ***Te quiero***

A la universidad que por mucho tiempo ha sido como mi segunda casa
UNIVERSIDAD AUTONOMA METROPOLITANA
UNIDAD IZTAPALAPA
A mi asesor de Tesis
Dr. GABRIEL ESCARELA PÉREZ
Gracias por darme la oportunidad de trabajar contigo y apoyarme en todo momento.

A mi amigo
Dr. JOHN NEWELL
Thank you for letting me visit the Biostatistics Unit of the Clinical Research Facility in NUI Galway and for continuing my research under your supervision, for supporting me and helping me feel at home.

Por último
CONACYT
Gracias por haberme brindado el apoyo económico para poder concluir mis estudios satisfactoriamente.

Índice general

Presentación

El estudio de datos de supervivencia juega un papel muy importante en la investigación médica, pues una comparación entre la supervivencia observada en dos grupos de pacientes puede llevar a validar un determinado tratamiento o alternativamente, a identificar un factor de riesgo significativo. En general, los estudios que intentan evaluar la supervivencia en una determinada situación presentan características particulares en cada caso. La situación más común corresponde al estudio en el cual la supervivencia del paciente se analiza a partir de un determinado instante de tiempo, en el que se interviene sobre él (administración de un tratamiento, intervención quirúrgica, etc.).

El estudio de datos de supervivencia implica el seguimiento de los individuos a lo largo del tiempo, pudiéndose producir una serie de situaciones que complican la caracterización de los mismos. Con ésto, la situación más favorable ocurre cuando se puede observar de manera exacta el tiempo T de aparición del suceso de interés (muerte, aparición de complicaciones post-operatorias, rechazo de un órgano trasplantado, etc.). Por otra parte, es habitual que algunos de los pacientes se pierdan a lo largo del seguimiento. Por ejemplo, puede ocurrir que un paciente trasplantado deje de acudir a la consulta, perdiéndose su rastro, a efectos de observar el suceso de interés. Se sabe que el suceso de interés no ocurrió durante el tiempo que se ha seguido al paciente, pero pasado este tiempo, no se tiene conocimiento de cuándo sucedió.

En la presente tesis se trabajará con una base de datos de pacientes diagnosticados con cáncer de próstata. Una vez que, a los pacientes se les ha diagnosticado con este tipo de cáncer, los médicos tienen que tomar una decisión sobre qué tratamiento se le tiene que dar para combatir dicho mal.

Estrategias conservadoras pueden resultar insuficientes para pacientes que buscan beneficiarse de un tratamiento mientras que para otros suelen ser estrategias adecuadas. Estrategias agresivas pueden resultar excesivas para algunos pacientes mientras que para otros implica la restauración de su salud (Stephenson, 2002). Actualmente, organizaciones tales como la American Urological Association y la National Comprehensive Cancer Network

han emitido guías para la elección del tratamiento de pacientes diagnosticados con cáncer de próstata basándose únicamente en la esperanza de vida del paciente, la cual es calculada a partir de datos poblacionales, ignorando por completo la información concerniente al estado de salud del paciente.

A pesar de que estas guías tienen buena aceptación en la práctica médica, la determinación de qué grupos de pacientes están lo suficientemente enfermos para beneficiarse de los distintos tipos de tratamientos no está bien definida, ya que los beneficios de los tratamientos tomados no se desarrollan hasta ocho o diez años después del tratamiento. Esta característica hace que el padecimiento sea difícil de tratar, y que los urólogos se queden en un dilema con la elección entre tratar agresivamente, tratar de manera conservadora, o simplemente aplazar el tratamiento y controlar la progresión de la enfermedad.

Por lo tanto, es importante desarrollar una herramienta predictiva que considere variables auxiliares que ayuden a decidir qué tipo de tratamiento es el más adecuado para cada paciente, de tal manera que se incrementen sus probabilidades de supervivencia, tomando en cuenta también las características de salud del paciente.

Objetivos

En esta investigación se pretende determinar un modelo estadístico para estimar la probabilidad de que un paciente fallezca dentro de un periodo arbitrario de tiempo de acuerdo al tipo de tratamiento y en presencia de variables auxiliares, las cuales deben incluir el estado de salud del paciente. Los parámetros del modelo podrán ser empleados para construir criterios que ayuden a decidir qué tipo de tratamiento es el más adecuado para incrementar las probabilidades de supervivencia de cada paciente.

En este estudio, se buscará emplear una metodología que permita incluir datos de pacientes cuyos registros presentan al menos una variable no observada.

Objetivos específicos:

- Emplear un método gráfico para visualizar las funciones empíricas de la incidencia.
- Desarrollar una metodología basada en verosimilitud para incluir variables explicativas incompletas.
- Proponer un modelo para el análisis de supervivencia que permita incluir variables explicativas tanto demográficas como aquellas que conciernen al estado de salud del paciente.
- Estudiar la forma funcional adecuada para cada variable explicativa continua que será introducida en el modelo propuesto.
- Proponer un criterio de decisión para la elección del mejor tratamiento para el paciente, en términos de los parámetros del modelo propuesto.

Capítulo 1

Cáncer de Próstata

Antes de comenzar con la parte matemática de este trabajo, en este capítulo se hablara de lo que es el cáncer de próstata, de las diferentes maneras que se presenta, de las causas que lo generan, de los tratamientos hacia esta enfermedad, de algunos términos médicos importantes y de la descripción de los datos con los que se trabajara posteriormente.

1.1. Definición

Es denominado cáncer de próstata al que se desarrolla en uno de los órganos glandulares en el sistema reproductor masculino llamado próstata.

Nuestro organismo está conformado principalmente por tejidos que a su vez está conformado por un conjunto de células, que se dividen de forma regular con el fin de reemplazar a las ya envejecidas o muertas y mantener la integridad y funcionamiento del organismo. La creación de estas nuevas células está regulada por mecanismos que indican cúando las células deben de comenzar a dividirse y cúando deben de quedarse estables.

Cuando estos mecanismos de creación de nuevas células se alteran comienzan los problemas ya que las células empiezan a dividirse incontrolablemente y con el tiempo dará lugar a un tumor.

Además, si estas células crecen sin control, adquieren la facultad de empezar a invadir los tejidos y órganos más cercanos además de proliferar en otras partes del organismo. Cuando comienzan a invadir los tejidos sanos más cercanos, se empiezan a implantar en los órganos, en nuestro caso los que están ubicados cerca de la próstata, dando origen al cáncer de próstata.

En la medicina los doctores han definido que el cáncer de próstata puede crecer de dos maneras:

- **Crecimiento local:** Se produce por crecimiento tumoral e invasión de

la cápsula prostática. Más tardíamente el tumor puede romper la misma y crecer invadiendo los tejidos y órganos periprostáticos. La invasión de la vejiga o el recto es tardía en el tiempo.

- **Diseminación hematógena:** Esta diseminación se realiza a través de los vasos sanguíneos, frecuentemente hacia el hueso.

1.2. Causas

Las causas principales del cáncer prostático que sugieren los médicos son:

- **Factores ambientales:** Quienes emigran de regiones de baja incidencia a regiones de alta incidencia mantienen una baja incidencia de cáncer prostático durante una generación y luego adoptan una incidencia intermedia. También se han identificado varios factores ambientales que podrían ser promotores del cáncer de próstata. Éstos incluyen:

  - Dieta alta en grasas animales.
  - La exposición a los gases del escape de los automóviles.
  - La polución del aire, cadmio, fertilizantes y sustancias químicas en las industrias de la goma, imprenta, pintura y naval.

- **Factores genéticos:** Aunque existen indicios que involucran a los factores genéticos en la causa del cáncer prostático, es difícil separar estos factores de los factores ambientales. Estudios genéticos han mostrado que existe un gen específico del cromosoma 1 ó gen HPC-1 que aumenta la probabilidad de contraer cáncer de próstata.

- **Agentes infecciosos :** Se ha considerado que los agentes infecciosos transmitidos por vía sexual podrían causar cáncer prostático, sin embargo, los estudios epidemiológicos han sugerido un aumento en el riesgo de cáncer prostático asociado con un mayor número de compañeros sexuales, una historia previa de enfermedad de transmisión sexual, frecuencia del acto sexual, relación con prostitutas y edad temprana de comienzo de la actividad sexual.

En contraste, otros estudios han sugerido que existe un mayor riesgo de cáncer prostático asociado con la represión de la actividad sexual, como un comienzo en edad más tardía, un pico más temprano y una cesación prematura de la actividad sexual. Por otra parte, algunos trabajos han mostrado

un mayor riesgo entre los pacientes que nunca estuvieron casados y un riesgo aún mayor entre aquellos que tuvieron niños, pero otros estudios no han mostrado una correlación significativa con el estado marital o con el número de hijos. De forma similar, los estudios de potenciales agentes infecciosos no han brindado resultados concluyentes, como tampoco proporcionan pruebas concretas para una causa infecciosa de cáncer prostático.

1.3. Tipos de cáncer de próstata

Aunque la próstata está formada por muchos tipos de células diferentes, más del 99 % de los cánceres de próstata se desarrollan sobre células glándulares.

- **Sarcomas.** Es una masa anormal de tejido maligno que se origina en un tejido conjuntivo, como pueden ser hueso, cartílago, grasa, músculo,vasos sanguíneos u otros. El término proviene de una palabra griega que significa crecimiento de la carne.

- **Adenocarcinoma.** El término médico del cáncer que se origina en las células glandulares se denomina adenocarcinoma. Los adenocarcinomas son un conjunto de cánceres muy frecuentes puesto que se originan en un tipo de células que se encuentran en continua división celular y que presentan mayor riesgo de mutaciones. Pueden presentarse inicialmente en forma de adenoma (un tumor glandular que es benigno).

Debido a que los otros tipos de cáncer de próstata son muy raros, cuando se habla de cáncer de próstata, lo más probable es que se refiera a un adenocarcinoma. Es raro encontrar sarcomas, carcinoma de células transicionales, de células pequeñas, epidermoides o escamosos. La próstata puede ser asiento de metástasis, de cáncer de vejiga, colón, pulmón u otras neoplasias.

1.4. Algunos términos médicos importantes

- **Carcinoma in situ.** El carcinoma in situ es una forma de cáncer con origen en células glandulares que no ha roto la capa basal y, por ello, no se ha extendido. El concepto tiene un interés especial ya que se considera que los cánceres in situ son susceptibles de ser curados con una simple extirpación tumoral.

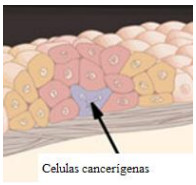

Figura 1.1: Ejemplo de carcinoma in situ

- **Anaplasia.** Se utiliza en medicina para describir la escasa diferenciación de las células que componen un tumor. Un tumor anaplásico es aquel cuyas células están poco diferenciadas o indiferenciadas, lo cual indica en general que su comportamiento es maligno, es decir tiene la capacidad de extenderse localmente a los tejidos vecinos y de diseminarse a otros órganos.

- **Microscopic focus o solo foci.** Se refiere cuando no hay un tamaño dado ya que no se puede ver, y la única forma de identificarlo es usando un microscopio.

- **Metástasis.** La metástasis es una teoría científica que supone la propagación del cáncer a un órgano distinto de aquel en que se inició. Ocurre generalmente por vía sanguínea. Aproximadamente, el 98 % de las muertes por cánceres no detectados, son debidas a la metastatización de éste. En realidad, aunque es la más conocida, la metástasis no se limita solo a la propagación de células cancerosas, sino que se habla de metástasis cuando un émbolo (masa sólida, líquida o gaseosa que se libera dentro de los vasos y es transportada por la sangre a un lugar del organismo distinto del punto de origen) desarrolla en el lugar donde produce la embolia el mismo proceso de su lugar de origen (cáncer, infecciones).

1.5. Tratamientos

- **Resección transuretral (RTU).** Es una intervención quirúrgica urológica que consiste en la extirpación de tejidos enfermos de uretra, próstata y vejiga accediendo a ellos a través de la luz uretral con un aparato endoscópico llamado resectoscopio.

- **Prostatectomía simple.** Es la cirugía para extirpar la parte interna de la glándula prostática a través de una incisión quirúrgica en la parte baja del abdomen con el fin de tratar el agrandamiento de la próstata.

- **Prostatectomia radical.** Es la cirugía para extirpar toda la glándula prostática, al igual que algunos tejidos que se encuentran alrededor de ésta, con el fin de tratar el cáncer de próstata.

1.6. Elección de tratamiento

En la actualidad los tratamientos para combatir el cáncer de próstata son multidisciplinarios, ya que distintas especialidades trabajan en conjunto para elegir la terapia adecuada y ofrecer una mayor probabilidad de curación. Muchos de los médicos recomiendan seguir un protocolo para elegir un adecuado tratamiento en contra del cáncer de próstata basándose en la experiencia científica.

Algunos factores que en la actualidad influyen en la elección adecuada en contra del cáncer de próstata son los siguiente:

- Estado de la enfermedad.
- Qué tan agresivas son las células cancerosas.
- Niveles de PSA (prostate-specific antigen) en el momento del diagnóstico.
- La edad y la esperanza de vida independientemente del cáncer de próstata.
- Preferencia del paciente

Probablemente el médico también tendrá en cuenta si además del cáncer de próstata, existen otras enfermedades importantes, que puedan dificultar la realización de algún tratamiento específico. El tratamiento propuesto por el especialista no va a ser el mismo en todos los pacientes.

1.7. Descripción de los datos

En la presente tesis se están considerando datos de 2687 hombres Estadounidenses diagnosticados con cáncer de próstata que fueron registrados en el US National Cancer Institute's Surveillance Epidemiology and End Results (SEER) durante 1988 y a los cuales se les dio seguimiento hasta el 1 de enero del 2008. La SEER es un programa encargado de la vigilancia epidemiológica, y consiste en registros de supervivencia de personas diagnosticadas con algún tipo de cáncer en los Estados Unidos. En 1973 se comenzó

a recabar información del cáncer de próstata y dicho registro se comenzó hacer en ciertas regiones geográficas, en las cuales estaba trabajando la SEER, este programa recoge información básica de los diagnósticos, el tratamiento y la mortalidad. Para pacientes diagnosticados con cáncer de próstata, los registros de la SEER tienen la siguiente información adicional: demografía, fecha del diagnóstico de cáncer, fecha del fallecimiento del paciente, causa del fallecimiento del paciente, tratamiento, grado del tumor, etapa y desde 1988 el tamaño del tumor. En esta base de datos solo fueron incluidos los casos de adenocarcinoma de próstata ya que es el tipo de cáncer más frecuente. Es raro encontrar sarcomas, carcinoma de células transicionales, de células pequeñas, epidermoides o escamosos por esta razón fueron excluidos. También fueron excluidos los casos diagnosticados por autopsia o por certificados de muerte.

Hay variables en este estudio que a simple vista se puede ver que son de suma importancia en la determinación de la gravedad y claro, de la supervivencia de los pacientes con cáncer de próstata, en el presente trabajo la etapa del tumor(en inglés *stage*) y el grado del tumor (en inglés *grade*) juegan un papel muy importante en lo antes mencionado, ya que cuentan con observaciones faltantes (en inglés *missing observations*). La etapa del tumor describe como el cáncer se va extendiendo, mientras que el grado describe la apariencia del cáncer e indica la rapidez con la que el cáncer está creciendo. En esta base hay un porcentaje de las variables etapa y grado del tumor con valor desconocido y por lo tanto se considera faltante o perdido. Aunque a ciencia cierta no se sabe el por qué de la existencia de datos faltantes, lo que si se puede notar es que juegan un papel muy importante en la determinación del pronóstico del paciente y la elección del tipo de tratamiento que se le dará, lo que se intuye es que el tratamiento y otros posibles factores pueden ayudar a predecir la etapa y el grado del tumor cuando estos son datos faltantes.

La respuesta de interés en el presente estudio es el tiempo en el que se diagnosticó al paciente hasta su fallecimiento por una de las causas siguientes:

1. Cáncer de próstata, que es la muerte por cáncer de próstata (en este caso por adenocarcinoma) que equivale al 16 % de los datos

2. Muerte por otras causas que equivale al 64 % de los datos

El otro 20 % de las observaciones son censuradas.

A continuación se muestra la descripción de las variables explicativas mencionadas anteriormente:

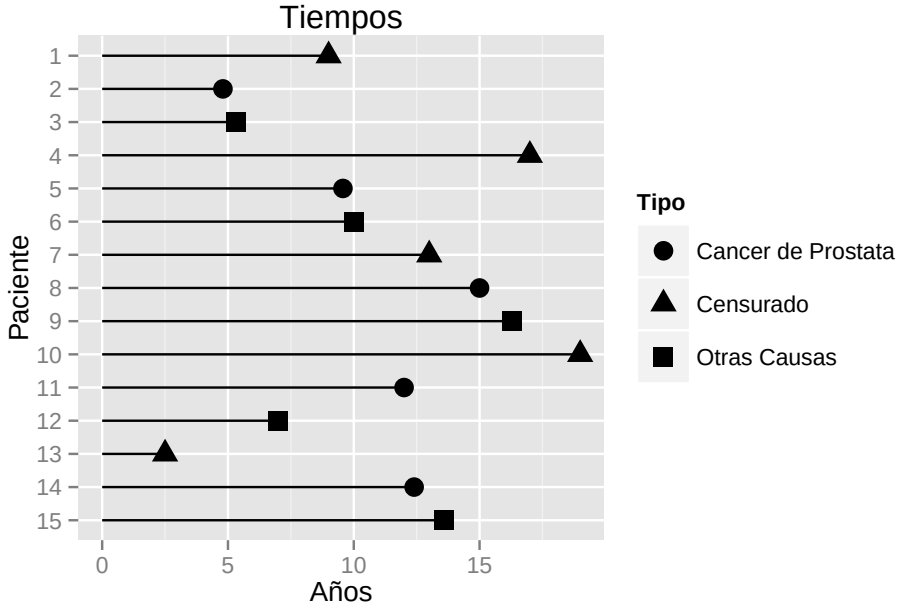

Figura 1.2: Ejemplo de los tiempos de acuerdo al status del paciente

- **Tratamiento.** Tipos de tratamiento, I no quirúrgico o quirúrgico pero no directamente por cáncer(18 %); II una cirugía simple ya sea la resección transuretral de la próstata, la prostatectomía crio, una escisión de la lesión o una simple prostatectomia (56 %) ; III cirugía radical(26 %).

- **Etapa.** Es la variable que se refiere a la etapa del cáncer y tiene las siguientes categorías: IS, carcinoma in situ (28 %);I localizado (17 %); II tempranamente avanzado (15 %); III tardemente avanzado (10 %); IV metástasis (6 %). El 25 % de los datos son faltantes.

- **Grado.** Es una variable categórica que se refiere al grado del tumor y toma los siguientes valores: I si el tumor es diferenciable (59 %); II si el tumor es moderadamente diferenciable(6 %); III si el tumor es pobremente diferenciable o si el tumor es anaplasia o no es diferenciable(ambos suman el 10 % de los datos) y datos faltantes (6 %).

- **Tamaño.** Esta variable se refiere al tamaño del tumor, en el cual podemos encontrar: no hay masa del tumor,no hay tumor encontrado y microscopic focus o solo foci (63 %) y mediciones continuas del tamaño que son mayores a 0 (37 %).

- **Edad.** Edad en años del paciente al momento del diagnóstico.

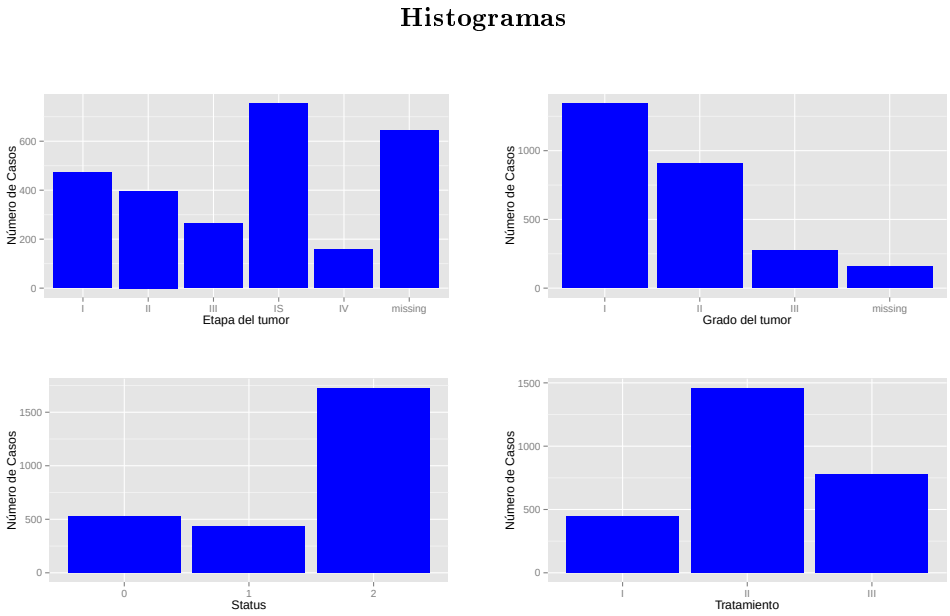

Figura 1.3: Histogramas de algunas variables que se midieron en el estudio

- **Raza.** Es la variable categórica que se refiere a la raza del paciente y toma los siguientes valores: I si es de raza blanca (88 %); II si es de raza negra (8 %) y III si es de otra raza (6 %).
- **Estado civil.** Esta variable se refiere a la situación civil del paciente. I si el estado es casado (79 %) y II otra estado civil (21 %).

Una vez que se ya se tiene noción de lo que es el cáncer de próstata y de la descripción de los datos que se trabajaran en esta tesis se procederá a desarrollar la parte matemática en los próximos capítulos para finalizar con la aplicación y los resultados.

Capítulo 2

Introducción

En este capítulo se dará una breve noción del análisis de supervivencia (punto de partida para analizar los datos descritos en el capítulo 1), de las distribuciones importantes para este tipo de análisis que en este caso son la función de supervivencia y la fuerza de mortalidad definidas como distribuciones de probabilidad, además de las dificultades que se tienen al trabajar con datos de supervivencia. Generalmente este análisis se trabaja sobre grupos de individuos de los cuales el interés se basa en la tasa de muerte que en este caso es la fuerza de mortalidad, por lo que también se presentara un modelo que permite comparar estos grupos en base a la tasa de muerte de cada uno de ellos.

2.1. Análisis de supervivencia de un sólo resultado

El análisis de supervivencia es una frase muy usada para describir el análisis de datos en forma de tiempos, en los cuales está bien definido el tiempo de origen hasta la ocurrencia de algún evento en particular o hasta el final del tiempo de seguimiento. En la investigación médica, el tiempo de origen corresponde a la inclusión de un individuo dentro de un estudio experimental, tal como lo es un ensayo clínico para comparar dos o más tratamientos. Esto puede coincidir con el diagnostico de una condición en particular, el inicio de un tratamiento, o una ocurrencia de algún evento adverso. Si al final del tiempo se encuentra con la muerte del paciente, los datos resultantes son literalmente *tiempos de supervivencia*.

En análisis de supervivencia el interés se centra sobre un grupo o grupos de individuos para quienes es definido un punto de fallo, que más tarde es llamado fracaso, este ocurre después de un cierto tiempo el cual es llamado

tiempo de fracaso. Como ejemplos de tiempos de fracaso podemos incluir el tiempo de vida de los componentes de una máquina en la industria, la duración de huelgas o periodos de desempleo en economía, el tiempo que puede tomar un sujeto en completar unas preguntas de un experimento psicológico y los tiempos de supervivencia de un paciente en un ensayo clínico. Para determinar precisamente un tiempo de fracaso hay tres requerimientos:

- El tiempo de origen debe de ser definido.
- La escala de medida para el tiempo debe de ser acordada.
- Debe de ser claro por qué sucedió el evento.

En el presente escrito se trabajará con aplicaciones médicas por lo que el evento de interés es la muerte, muerte de una causa especifica, la primera recaída de una enfermedad después de un tratamiento, o la aparición de una nueva enfermedad.

Características especiales de los datos de supervivencia

Es importante considerar las razones por la que los datos de supervivencia no son compatibles con procedimientos estadísticos estándares. En primer lugar se tiene que los datos de supervivencia no son en general distribuidos simétricamente. Típicamente, un histograma con los datos de supervivencia de un grupo de individuos similares tiende a ser de "cola larga" a la derecha del intervalo que contiene la mayoría de las observaciones. En consecuencia, no es razonable suponer que los datos proviene de una distribución normal. Esta dificultad, se podría resolver aplicando una transformación a los datos para obtener una distribución simétrica, por ejemplo aplicando un logaritmo natural. Sin embargo, una metodología más satisfactoria es el adoptar un modelo alternativo cuya distribución sea más apropiada para la bondad de ajuste de los datos.

Una segunda dificultad en el análisis de los datos de supervivencia es que los tiempos de supervivencia frecuentemente están censurados lo cual hace a los métodos estándares poco apropiados. El tiempo de un individuo se dice censurado cuando el evento de interes no ha sido observado. Esto puede ocurrir cuando al final del periodo de estudio varios individuos siguen vivos, que es el caso de la aplicación a la medicina. Un ejemplo claro de esto es que algunos pacientes que mueren por enfermedad en el corazón, no pueden morir por cáncer de pulmón. Otra forma de censura es, cuando el estatus de supervivencia de un individuo en el estudio no es bien sabido, porque se le ha perdido de vista ya sea porque tal vez el individuo después de ser

reclutado para el estudio, se muda a otro país o simplemente no se le puede encontrar. La única información disponible acerca de la supervivencia del individuo censurado es la última fecha en que se le vio con vida.

Un individuo que entra en un estudio en el tiempo t_0 muere en tiempo $t_0 + t$. Sin embrago, si el tiempo de supervivencia correspondiente es censurado entonces t es desconocido, ya sea por que el individuo sigue vivo o porque se le ha perdido de vista. Si el individuo fue visto por última vez en el tiempo $t_0 + c$, el tiempo c es conocido como el tiempo de supervivencia censurado por la derecha.

Otra forma de censura es la llamada censura por la izquierda que sucede cuando el tiempo de supervivencia t (un valor desconocido) es menor que un cierto valor t_1 (conocido). Para ilustrar este tipo de censura supongamos que nuestro interés se centra en la recaída de algún tipo de cáncer en particular, en el cual los individuos con este padecimiento fueron operados para remover un tumor cancerígeno. Tres meses después de su operación, los pacientes fueron examinados para determinar si recayeron por cáncer. A ese tiempo, algunos pacientes recayeron por cáncer. Para estos pacientes, el tiempo de recaída es menor a tres meses, por lo que sus tiempos de recaída son censurados por la izquierda, ya que no se conoce con exactitud el tiempo, lo único que se sabe es que fue menor a tres meses.

Algunos ejemplos

Supervivencia de pacientes con mieloma múltiple

La mieloma múltiple es una enfermedad maligna caracterizada por la acumulación de células anormales en la medula ósea. El desarrollo de estas células anormales dentro del hueso puede causar dolor y destrucción en el tejido óseo. Pacientes con esta enfermedad también experimentan anemia, recurrentes infecciones y debilidad. El objetivo de este ejemplo, fue examinar la relación entre los valores de ciertas variables explicativas y los tiempos de supervivencia de los pacientes. En el estudio, la principal variable de respuesta fue el tiempo, en meses, del diagnostico hasta la muerte por mieloma múltiple.

Los datos fueron obtenidos de 48 pacientes, todos ellos de edades entre 50 y 80 años. Algunos de estos pacientes no estaban muertos al tiempo en el que se realizó el estudio, así estos individuos contribuyeron a tiempos de supervivencia censurados por la derecha. El estatus de supervivencia de un individuo es cero si la observación es censurada y uno si murió por mieloma múltiple. Al momento del diagnostico, los valores de las variables explicativas fueron recabadas para cada paciente. Estos incluyen la edad del paciente en

años, su sexo (1=masculino,2=femenino), niveles de nitrógeno en la sangre (BUN), calcio(Ca) y hemoglobina (Hb), el porcentaje de plasmas en la médula ósea (Pcells) y el indicador de la variable (Protein) que indica si tiene o no la proteína Bence-Jones en la orina (1=presenta,0=no presenta).

Comparación de dos tratamientos de cáncer de próstata

Un estudio clínico controlado aleatoriamente para comparar dos tratamientos de cáncer fue realizado por el grupo de investigación de Administration Cooperative Urological. Los dos tratamientos fueron realizados, uno aplicando un placebo y el otro aplicando 1.0 mg de diethylstilbestrol (DES). El tratamiento fue administrado diariamente por un mes. El tiempo de origen del estudio es el día en el que el paciente comenzó con su tratamiento, y el tiempo final es el día en que murió de cáncer de próstata. El estudio incluyo otro número de factores que pudieron influir en la enfermedad, estos son: la edad del paciente cuando entra al estudio, el nivel de hemoglobina en la sangre medida en $mg/100\ ml$, el tamaño del tumor en cm^2 y el índice de Gleason, que indica qué tan avanzado está el tumor.

En este caso los tiempos de supervivencia de los pacientes que murieron por otras causas o quienes se perdieron durante el tiempo del estudio fueron censurados. De la misma manera se consideró una variable para el estado del paciente, la cual vale uno si el paciente murió por cáncer de próstata, y cero si el tiempo de supervivencia es censurado por la derecha. La variable asociada con el tratamiento toma el valor de 2 si al individuo se le aplica el tratamiento de DES y uno si al individuo se le aplicó el tratamiento de placebo. El objetivo de este estudio es determinar cómo afectan los tratamientos de DES y de placebo a los tiempos de supervivencia.

2.2. Distribución de los tiempos de fracaso

Para comenzar la introducción del modelado estadístico de datos de supervivencia, es pertinente considerar unas características relevantes de las distribuciones de probabilidad para el análisis; para esto, se partirá de que población es homogénea y que cada individuo tiene un "tiempo de fracaso" que en este estudio significa la muerte. A continuación se examinan algunas especificaciones de una variable aleatoria positiva T, la cual está asociada con el tiempo para que ocurra el evento, y después se consideran varias distribuciones especiales que son de utilidad para el ajuste de los datos correspondientes.

En los datos de supervivencia hay dos funciones que son muy importantes, la *función de supervivencia* y la *fuerza de mortalidad* (también conocida en inglés como *hazard function*).

El tiempo actual de supervivencia de un individuo, t, puede ser considerado como el valor de una variable T, que toma valores no negativos. Los diferentes valores que puede tomar T tienen una *distribución de probabilidad*, y así se dice que T es la variable aleatoria asociada con el tiempo de supervivencia. Ahora supóngase que T tiene una distribución de probabilidad con una *función de densidad de probabilidad* $f(t)$. La función de distribución de T está dada por:

$$F(t) = P\{T < t\} = \int_0^t f(u)\,du \tag{2.1}$$

y representa la probabilidad de que el tiempo de supervivencia sea menor que el valor t.

La función de supervivencia, $S(t)$, se define como la probabilidad de que el tiempo de supervivencia sea mayor o igual a t

$$S(t) = P\{T \geq t\} = 1 - F(t) \tag{2.2}$$

La función de supervivencia puede ser usada para representar la probabilidad de que un individuo sobreviva desde el tiempo de su origen hasta más allá del tiempo t.

La fuerza de mortalidad es extensamente usada para expresar el riesgo o la tasa de muerte al tiempo t, y de esta forma se tiene la probabilidad de que un individuo muera al tiempo t, condicionado a que ha sobrevivido hasta este tiempo. Para una definición formal de la fuerza de mortalidad, considérese la probabilidad de la variable aleatoria asociada con el tiempo de supervivencia de un individuo, T, que vive entre t y t+ Δ_t, condicionado a que T es mayor o igual a t, y la cual se escribe de la siguiente forma $P\{t \leq T < t + \Delta_t | T \geq t\}$. Esta probabilidad condicional entonces puede ser

expresada como la probabilidad unitaria de tiempo dividida por el intervalo de tiempo Δ_t. La fuerza de mortalidad $h(t)$, es entonces el valor límite de esta cantidad, cuando Δ_t tiende a cero, así

$$h(t) = \lim_{\Delta_t \to 0} \left\{ \frac{P\{t \leq T < t + \Delta_t | T \geq t\}}{\Delta_t} \right\} \tag{2.3}$$

La función $h(t)$ es también referida como la *tasa de mortalidad*, la *tasa de muerte instantánea*, la *tasa de intensidad*, la *fuerza de mortalidad*, o simplemente la *tasa de muerte*.

De la ecuación (2.3), $h(t)\Delta_t$ es aproximadamente la probabilidad de que un individuo muera en el intervalo $(t, t + \Delta_t)$, condicionada a que la persona a sobrevivido al tiempo t. Por ejemplo si el tiempo de supervivencia es medido en días, $h(t)$ es aproximadamente la probabilidad de que un individuo, quien está vivo hasta el tiempo t, muera en los siguientes días. Por esta razón la fuerza de mortalidad es interpretada como el riesgo de muerte al tiempo t.

De la definición de fuerza de mortalidad en la ecuación (2.3), se puede obtener una relación entre la función de supervivencia y la fuerza de mortalidad. Acordado con los resultados estándares de la teoría de probabilidad, la probabilidad de que un evento A, condicionado a que ocurra un evento B, esta da por $P\{A/B\} = P\{AB\}/P\{B\}$, donde $P\{AB\}$ es la probabilidad de que suceda A y B. Usando este resultado, la probabilidad condicional en la definición de la fuerza de mortalidad en la ecuación (2.3) es

$$\frac{P\{t \leq T < t + \Delta_t\}}{P\{T \geq t\}}$$

$$= \frac{F(t + \Delta_t) - F(t)}{S(t)}$$

donde $F(t)$ es la función de distribución de T. Entonces,

$$h(t) = \lim_{\Delta_t \to 0} \left\{ \frac{F(t + \Delta_t) - F(t)}{\Delta_t} \right\} \frac{1}{S(t)}$$

$$= \{F'(t)\} \frac{1}{S(t)}$$

$$= \frac{f(t)}{S(t)} \tag{2.4}$$

Otra forma de escribir la tasa de muerte de acuerdo a la ecuación anterior es la siguiente:

$$h(t) = -\frac{\mathrm{d}}{\mathrm{dt}} [\log S(t)] \tag{2.5}$$

y se obtiene que

$$S(t) = \exp[-H(t)] \tag{2.6}$$

donde

$$H(t) = \int_0^t h(u)\,du \tag{2.7}$$

La función $H(t)$ es muy característica en el análisis de supervivencia, y es llamada la *fuerza de mortalidad integrada* (también conocida en ingles como *comulative hazard*).

Una función $h(x)$ es una fuerza de mortalidad sí y solo si satisface las siguientes propiedades:

$$h(x) \geq 0, \forall x.$$

$$\int_0^\infty h(u)\,du = \infty$$

Estas propiedades son necesarias ya que

$$h(x) = \frac{f(x)}{S(x)} \geq 0$$

y

$$\int_0^\infty h(x)\,dx = \int_0^\infty -d[\log S(x)]\,dx = -\log S(x)|_0^\infty = \infty$$

Estas propiedades son suficientes ya que la función de distribución resultante $F(x)$ es válida; que es, en términos de la fuerza de mortalidad $h(x)$:

$$F(-\infty) = F(0) = 1 - \exp[-\int_0^0 h(t)\,dt] = 0$$

y

$$F(\infty) = 1 - \exp[-\int_0^\infty h(t)\,dt] = 1$$

y $F(x)$ es una función creciente de x ya que $\int_0^x h(t)\,dt$ es una función creciente de x.

De la ecuación (2.6), la fuerza de mortalidad integrada puede ser obtenida de la función de supervivencia, así:

$$H(t) = -\log[S(t)] \tag{2.8}$$

En el análisis de datos de supervivencia, la función de supervivencia y la fuerza de mortalidad son las cantidades más relevantes a estimar.

2.3. Algunas distribuciones Especiales

En el presente trabajo se consideran algunas distribuciones especiales que son utilizadas en los datos de supervivencia. Por supuesto aquí sólo son tomadas en cuenta aquellas variables que toman valores no negativos que son las principales candidatas en este estudio.

2.3.1. Distribución Weibull

Se dice que T tiene distribución Weibull con parámetro de escala λ y parámetro de forma α, y se denota $T \sim \text{WEI}(1/\lambda, \alpha)$, cuya función de densidad está dada por

$$f(t) = \alpha\lambda^{\alpha} t^{\alpha-1} \exp[-(\lambda t)^{\alpha}], \alpha, \lambda > 0.$$

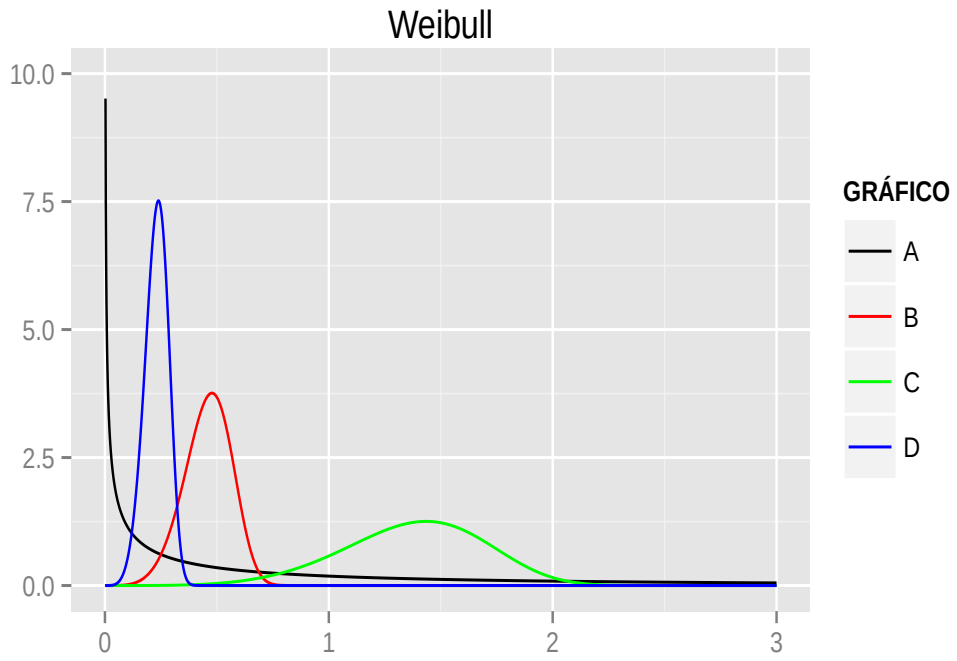

Figura 2.1: Función de densidad de probabilidad para distintos valores de los parámetros forma y escala.

La fuerza de mortalidad, la función de supervivencia y la fuerza de mortalidad integrada de la distribución Weibull son:

$$h(t) = \alpha\lambda^{\alpha} t^{\alpha-1},$$

$$S(t) = \exp[-(\lambda t)^{\alpha}],$$

y respectivamente

$$H(t) = (\lambda t)^{\alpha},$$

La distribución Weibull cuenta con propiedades muy interesantes ya que es un modelo muy flexible en sus parámetros, en particular el de forma. El parámetro de forma α juega un rol muy importante en la forma de cómo se ve la distribución Weibull. Por ejemplo $0 < \alpha < 1$, la función de densidad de probabilidad tiende a infinito conforme el tiempo tiende a cero y decrece rápidamente a cero conforme se incrementa el tiempo, y de la misma manera la fuerza de mortalidad tiende a cero conforme aumenta el tiempo. Cuando $\alpha = 1$, la distribución Weibull se reduce a la distribución exponencial con fuerza de mortalidad constante λ. Para $\alpha > 1$, la función de densidad de probabilidad comienza en cero y se incrementa hasta alcanzar un máximo de $(1/\lambda)(1-(1/\alpha))^{1/\alpha}$ para después decrecer a cero conforme se incrementa el tiempo, mientras que la fuerza de mortalidad aumenta conforme pasa el tiempo.

2.3.2. Distribución Pareto

La distribución Pareto se propuso por primera vez como un modelo para la distribución de ingresos. También se utiliza como modelo para la distribución de las poblaciones dentro de una zona determinada. La densidad Pareto se define como:

$$f(t) = \frac{\zeta}{\omega}(\frac{\omega}{\omega + t})^{\zeta+1}, (\zeta > 0, \omega > 0)$$

La tasa de mortalidad, la función de supervivencia y la fuerza de mortalidad integradade est densidad son:

$$h(t) = \frac{\zeta}{\omega + t}$$

$$S(t) = (\frac{\omega}{\omega + t})^{\zeta}$$

$$H(t) = -\zeta \log(\frac{\omega}{\omega + t})$$

2.3.3. Distribución Log-Logística

La distribución log-logística es utilizada en los casos donde la fuerza de mortalidad tiene diferentes cambios, a diferencia de la distribución Weibull que cuenta con algunas limitaciones por que su función de mortalidad es

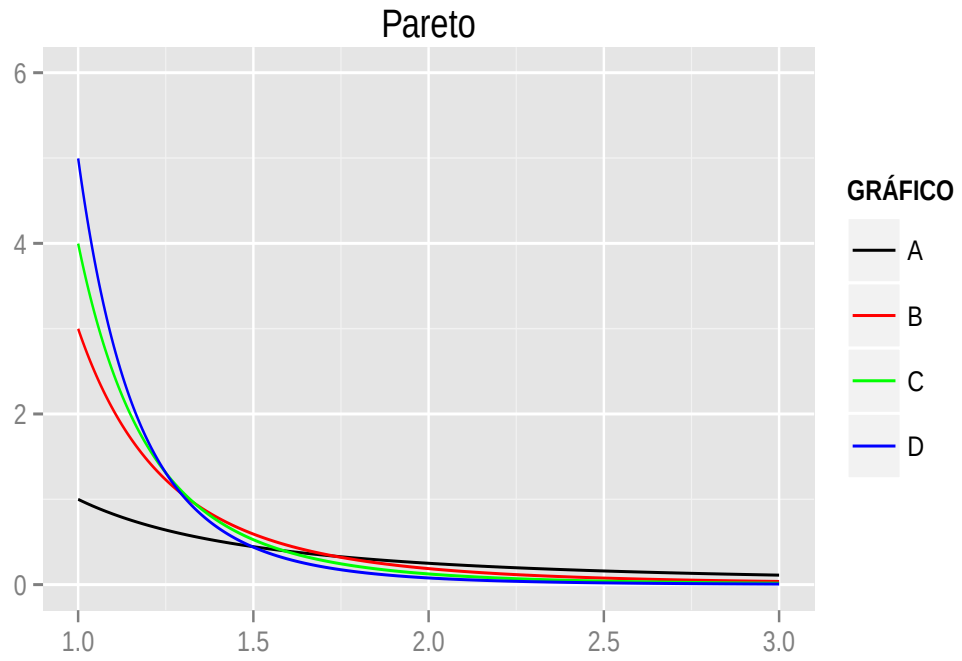

Figura 2.2: Función de densidad de probabilidad para distintos valores de sus parámetros.

monótona en el tiempo. Por ejemplo supongamos que tenemos un trasplante de corazón a un paciente, en este caso la tasa de muerte se incrementa en los primeros 10 días después de la operación, mientras el cuerpo acepta el nuevo órgano. Para que después la tasa decrezca conforme el paciente se recupera. En estas situaciones una forma particular de una función de mortalidad es la siguiente:

$$h(t) = \frac{e^{\theta}\kappa t^{\kappa-1}}{1 + e^{\theta}t^{\kappa}}$$

para $0 \leq t < \infty$, $\kappa > 0$. La función de supervivencia de la fuerza de mortalidad anterior está dada por

$$S(t) = [1 + e^{\theta}t^{\kappa}]^{-1}$$

y la función de densidad de probabilidad es

$$f(t) = \frac{e^{\theta}\kappa t^{\kappa-1}}{(1 + e^{\theta}t^{\kappa})^2}$$

Esta es la densidad de una variable aleatoria T con distribución *log logística*, con parámetros θ y κ. La distribución es llamada así por que la variable $logT$ tiene distribución logística, y es una función de densidad de probabilidad simétrica muy similar a la distribución normal.

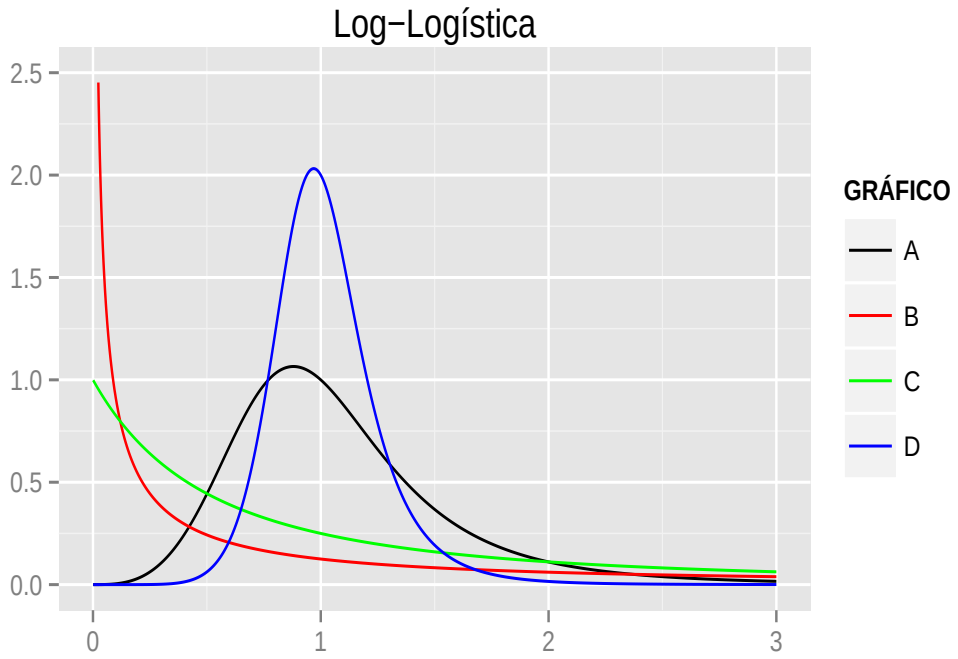

Figura 2.3: Función de densidad de probabilidad para distintos valores de sus parámetros.

2.3.4. Distribución Log Normal

Se dice que una variable aleatoria T tiene una distribución log normal con parámetros μ y σ, si $\log T$ tiene una distribución normal con media μ y varianza σ^2. La función de densidad de probabilidad de T está dada por

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} t^{-1} exp[-(\log t - \mu)^2/2\sigma^2]$$

para $0 \leq t < \infty$, $\sigma > 0$. De la cual la función de supervivencia puede ser derivada

$$S(t) = 1 - \phi(\frac{\log t - \mu}{\sigma})$$

donde $\phi(.)$ es la función de distribución normal estándar dada por

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp[-\frac{u^2}{2}]\, du$$

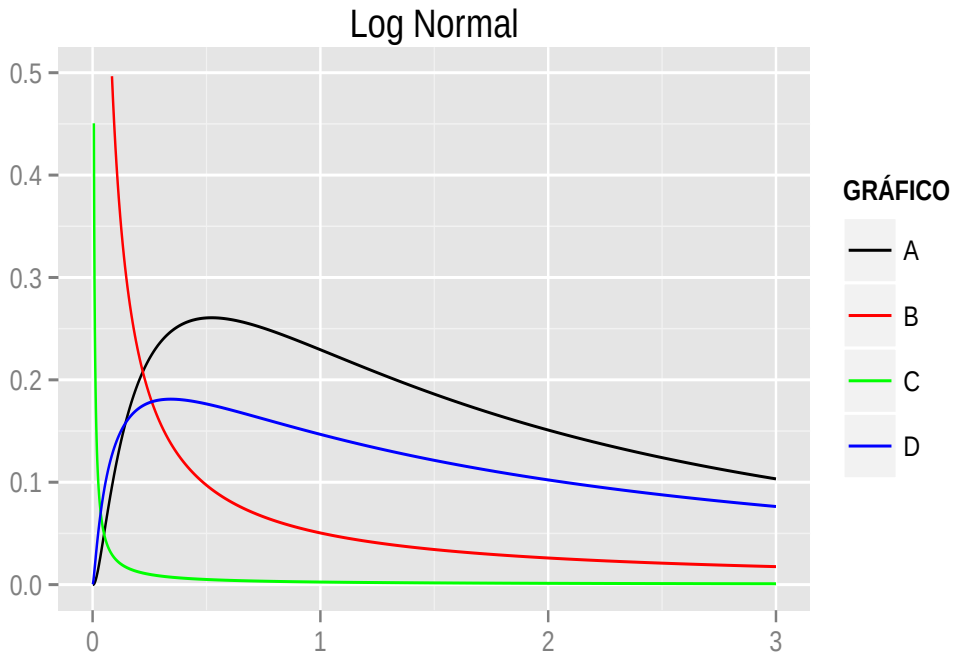

Figura 2.4: Función de densidad de probabilidad para distintos valores de sus parámetros.

2.4. Modelos de Regresión

A través del modelado de los datos de supervivencia, se puede obtener que la supervivencia de los pacientes en un grupo depende de una o más variables explicativas, cuya información es recabada para cada paciente desde su inclusión al evento (tiempo de origen). Por ejemplo en el estudio con Mieloma múltiple dado en el Ejemplo 1, el objetivo es determinar cuál de las siete variables explicativas tiene un gran impacto sobre los tiempos de supervivencia de los pacientes. En el Ejemplo 2 se tienen los tiempos de supervivencia de los pacientes a dos tratamientos de cáncer de próstata, el objetivo primario es identificar si los pacientes en los dos grupos de tratamiento tienen diferentes tiempos de supervivencia, porque variables adicionales como la edad del paciente y el tamaño del tumor pueden influenciar mucho en estos tiempos, por eso es importante tomar en cuenta esta variables cuando se evalúa alguno de estos tratamientos.

En el análisis de datos de supervivencia el interés central actúa sobre el riesgo o la tasa de muerte en algún tiempo posterior al tiempo de origen del estudio. Como consecuencia, una función importante en este modelado es la *fuerza de mortalidad*. Los modelos resultantes se forman a partir de modelos lineales encontrados en el análisis de regresión y en el análisis de datos de

los experimentos, donde la dependencia de ciertas variables explicativas es modelada.

Hay dos razones para la modelación de los datos de supervivencia, la primera es poder determinar cuál combinación de variables explicativas afecta la forma de la *fuerza de mortalidad*. En particular, el efecto que tiene el tratamiento sobre la tasa de muerte del objeto de estudio, así como el grado en que afecta cada variable explicativa a la fuerza de mortalidad. Otra de las razones por la cual se decide modelar los datos de supervivencia es poder obtener una estimación de la fuerza de mortalidad para el paciente, así se puede encontrar una estimación de la función de supervivencia, ya que se sabe de la relación que esta tiene con la fuerza de mortalidad.

Un modelo básico para los datos de supervivencia es el *modelo de riesgos proporcionales* (conocido en inglés como *proportional hazard model*). Este modelo fue propuesto por Cox (1972) y también es conocido como el *modelo de regresión de Cox*. El modelo se basa en la suposición de 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. A continuación se detallaran los modelos basados en esta suposición.

2.5. Un modelo para la comparación de dos muestras

Supóngase que se tiene una muestra aleatoria de pacientes clasificados en dos grupos independientes para recibir un tratamiento nuevo o un tratamiento estándar para combatir alguna enfermedad, y sean $h_s(t)$ y $h_N(t)$ las tasas de muerte al tiempo t para los pacientes de el tratamiento estándar y del nuevo tratamiento, respectivamente. Un modelo simple para los tiempos de supervivencia de dos grupos de pacientes, es suponer que las tasa de muerte al tiempo t para los pacientes sobre el nuevo tratamiento es proporcional a la tasa de muerte al mismo tiempo para los pacientes del tratamiento estándar. A este se le conoce como el *modelo de riesgos proporcionales* y puede ser expresado de la siguiente manera

$$h_N(t) = \psi h_s(t), \tag{2.9}$$

para valores no negativos de t, donde ψ es una constante. Una implicación de esta suposición es que las funciones de supervivencia correspondientes a los individuos para los tratamientos no se cruzan. Notemos que ψ es la razón de las tasas de muerte en algún tiempo para un individuo sobre el nuevo tratamiento relativo a uno sobre el tratamiento estándar, de esta manera ψ

es conocida como la *razón de riesgo* (en inglés *hazard ratio*). Para $\psi<1$, la tasa de muerte al tiempo t es más pequeña para un individuo sobre el nuevo tratamiento relativo a un individuo sobre el tratamiento estándar, de esto podemos concluir que el nuevo tratamiento es mejor que el estándar. Por otra parte si $\psi>1$, la tasas de muerte al tiempo t es más grande para un individuo sobre el nuevo tratamiento relativo al tratamiento estándar, por lo que podemos concluir que el tratamiento estándar es mejor que el nuevo tratamiento.

Ahora se tratará de generalizar la expresión en la ecuación (2.9) . Supóngase que se tienen disponibles los tiempos de supervivencia para n individuos y denótese la tasa de muerte del i-ésimo individuo como $h_i(t)$, $i = 1, 2, \ldots, n$. También se escribirá $h_s(t)$ como la tasa de muerte de un paciente sobre el tratamiento estándar, así la tasa de muerte para un paciente sobre el nuevo tratamiento es $\psi h_s(t)$. Dado que ψ no puede ser negativa, supóngase que $\psi = exp(\beta)$, nótese que los valores positivos de β se obtienen cuando la razón de riesgo ψ es más grande que la unidad, esto es cuando el nuevo tratamiento es mejor que el estándar.

Ahora sea X la variable indicadora, la cual toma valores de cero si al paciente se le aplica el tratamiento estándar, y uno si se le aplica el nuevo tratamiento. Si x_i es el valor de X para el i-ésimo paciente en el estudio, $i = 1, 2, \ldots, n$, la tasa de muerte para este paciente se puede escribir de la siguiente manera

$$h_i(t) = (exp(\beta x_i))h_s(t) \tag{2.10}$$

donde $x_i = 1$ si al i-ésimo paciente se le aplica el nuevo tratamiento y $x_i = 0$ de otro modo. El modelo mencionado anteriormente se le conoce como *modelo de riesgos proporcionales* para la comparación de dos grupos.

2.6. Modelo de riesgos proporcionales generalizado

Ahora se generalizará el modelo anterior y para ello supóngase que la tasa de muerte para algún tiempo en particular depende de los valores $x_1, x_2, \ldots, x_p$ de p variables explicativas, $X_1, X_2, \ldots, X_p$. Este conjunto de variables en el modelo de riesgos proporcionales puede ser representado por el vector $\mathbf{x}$, asi $\mathbf{x}^T = (x_1, x_2, \ldots, x_p)$. Sea $h_0(t)$ la tasa de muerte de referencia que corresponde a la del individuo cuya forma funcional de las variables explicativas ligadas a la constante ψ dan cero. La función $h_0(t)$ es llamada la *fuerza de mortalidad inicial* (en inglés, *baseline hazard function*). La tasa de muerte

para el i-ésimo individuo se puede escribir como:

$$h_i(t) = \psi(\mathbf{x}_i)h_0(t) \tag{2.11}$$

donde $\psi(\mathbf{x}_i)$ es una función de valores del vector de variables explicativas para el i-ésimo individuo. La función ψ puede ser interpretada como la mortalidad al tiempo t para el individuo que tiene variables explicativas $\mathbf{x}_i$, relativo a la mortalidad de un individuo que tiene vector $\mathbf{x}$=$\mathbf{0}$.

Como la razón de riesgo, $\psi(\mathbf{x}_i)$, no puede ser negativa, es conveniente que se defina como $\psi(\mathbf{x}_i) = exp(\eta_i)$ donde η_i es la combinacion lineal de las p variables explicativas en $\mathbf{x}_i$, esdecir

$$\eta_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$$

$$\eta_i = \sum_{j=1}^{n} \beta_j x_{ji}$$

En forma matricial, $\eta_i = \beta' x_i$, donde β es un vector con los coeficientes de las variables explicativas $x_1, x_2, \ldots, x_p$ en el modelo. El valor η_i es llamado la *componente lineal* del modelo, aunque también es conocido como el *índice de pronósticos* para el i-ésimo individuo. De esta manera el modelo de riesgos proporcionales generalizado puede ser expresado como

$$h_i(t) = (exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}))h_0(t). \tag{2.12}$$

Así este modelo puede re-expresarse de la siguiente forma

$$\log\left[\frac{h_i(t)}{h_0(t)}\right] = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}. \tag{2.13}$$

el modelo de riesgos proporcionales también puede ser visto como un modelo lineal del logaritmo de la razón de riesgo. Hay otras posibles formas para $\psi(x_i)$, aunque $\psi(x_i) = exp(\beta' x_i)$ es el que comúnmente se usa en los modelos de análisis de supervivencia.

Obsérvese que en el componente lineal no existe el término constante β_0. Si se agregara el término constante β_0, la tasa de muerte inicial se re-escalaria al dividir $h_0(t)$ por $exp(\beta_0)$, y el término constante se cancelaria, por lo que β_0 es redundante.

Una forma de evaluar el modelo de riesgos proporcionales dada la ecuación (2.12) es conocer la estimación de los coeficientes $\beta_1, \beta_2, \ldots, \beta_p$ de las variables explicativas en el componente lineal del modelo. Ya que es muy importante los efectos que puedan tener las p variables explicativas sobre la razón de riesgo y en algunos casos no es necesaria la estimación de $h_0(t)$, aunque existen

algunas formas de estimarla. Un método para estimar los parámetros del modelo (2.13) es usar el método de *máxima verosimilitud*. De esta manera, supóngase que se tienen disponibles datos de n individuos, de los cuales hay r distintos tiempos de muerte y $n-r$ tiempos de muerte censurados. Así sean $t_{(1)}, < t_{(2)} < \ldots < t_{(r)}$ los r tiempos de muerte ordenados, de manera que $t_{(j)}$ es el j-ésimo tiempo de muerte ordenado. El conjunto de individuos que se encuentran en riesgo al tiempo $t_{(j)}$ se denotará como $R(t_{(j)})$, así que $R(t_{(j)})$ es el conjunto de los individuos que se encuentran con vida y sin censura antes de $t_{(j)}$. La cantidad $R(t_{(j)})$ es conocida como el *conjunto de riesgo*.

Cox (1972) demostró que la función relevante de verosimilitud para el modelo de riesgos proporcionales está dada por

$$L(\beta) = \prod_{j=1}^{r} \frac{e^{x_j^{'}\beta}}{\sum_{\iota \epsilon R(t_{(j)})} e^{x_\iota^{'}\beta}}$$

donde x_j es el vector de variables explicativas para el individuo que muere en el j-ésimo tiempo de muerte ordenado $t_{(j)}$. La suma del denominador de la función de verosimilitud corresponde a los valores de $e^{x^{'}\beta}$ de todos los individuos que se encuentran en riesgo en el tiempo $t_{(j)}$. Se puede notar que el producto se toma de los individuos para los cuales se observó una muerte. Los individuos con tiempos censurados no tienen contribución en el numerador, su contribución se encuentra en la suma de los individuos en riesgo del denominador. La aproximación a la función de verosimilitud anterior no es la única que existe, pero en este caso sirvió para ejemplificar la estimación de los parámetros.

2.7. Modelos de riesgos proporcionales paramétricos

Cuando se utiliza el modelo de regresión de Cox para el análisis de los datos de supervivencia, no necesariamente se tiene que asumir una forma particular de la función de distribución de los tiempos de supervivencia, pero si se hace la suposición de una función de distribución en particular también es válido, así estos últimos son conocidos como *modelos paramétricos*

2.7.1. Evaluación de un modelo paramétrico para una muestra

Los modelos paramétricos son utilizados para evaluar conjuntos de observaciones de datos de supervivencia usando el método de máxima verosi-

militud. Así que primero considérese la situación donde los tiempos de supervivencia de n individuos no son observaciones censuradas, de esta forma, si la función de densidad de probabilidad de la variable aleatoria asociada al tiempo de supervivencia es $f(t)$, la verosimilitud de las n observaciones $t_1, t_2, \ldots, t_n$ es el siguiente producto

$$\prod_{i=1}^{n} f(t_i)$$

Esta verosimilitud es una función con parámetros desconocidos en la función de densidad de probabilidad, y la estimación de la máxima verosimilitud de estos parámetros nos da el máximo de la función de verosimilitud. En general es más práctico trabajar con el logaritmo de la función de verosimilitud, ya que los valores de los parámetros desconocidos en la función de densidad que maximizan el logaritmo de la verosimilitud son también los valores que maximizan a la verosimilitud.

Ahora considérese una situación un poco más real donde los datos de supervivencia incluyen uno o más tiempos de supervivencia. De la misma manera que anteriormente, supóngase que se tienen r tiempos de muerte $t_1, t_2, \ldots, t_n$ de n individuos y $n - r$ tiempos de supervivencia $t_1^*, t_2^*, \ldots, t_n^*$ son censurados por la derecha. De esta manera los r tiempos de muerte contribuyen de la siguiente forma

$$\prod_{j=1}^{r} f(t_j)$$

a la función de verosimilitud. Naturalmente no se puede ignorar la información acerca de la supervivencia de los $n - r$ tiempos de supervivencia de los individuos censurados. Si el tiempo de supervivencia es censurado al tiempo t^*, se dice que el tiempo de vida del individuo es al menos t^* y que la probabilidad de este evento es $P\{T \geq t^*\}$, que es lo que conoce como $S(t^*)$. Así cada observación censurada contribuye a la verosimilitud de esta forma. Por lo tanto la función de verosimilitud total queda expresada

$$\prod_{j=1}^{r} f(t_j) \prod_{\iota=1}^{n-r} S(t_j^*) \tag{2.14}$$

en el cual el primer producto es tomado sobre las r muertes y el segundo término es tomado sobre los $n - r$ tiempos de supervivencia censurados.

Una manera más compacta es suponer que los datos estan acomodados en n parejas de observaciones, donde el i-ésimo individuo es (t_i, δ_i), $i =$

$1, 2, \ldots, n$. En esta notación δ es una variable indicadora que toma el valor cero si el tiempo de supervivencia t_i es censurado y uno cuando t_i es un tiempo de supervivencia no censurado.

La función de verosimilitud se puede escribir como

$$\prod_{j=1}^{n} [f(t_j)]^{\delta_i} [S(t_j)]^{1-\delta_i} \tag{2.15}$$

Esta función es equivalente a la función mostrada en la expresión (2.14), y puede ser maximizada con respecto a los parámetros desconocidos en las funciones de densidad y de supervivencia. Agrupando de una manera adecuada la expresión (2.15) se puede obtener lo siguiente

$$\prod_{i=1}^{n} \left[\frac{f(t_i)}{S(t_i)}\right]^{\delta_i} [S(t_i)] \tag{2.16}$$

y de la ecuación (2.4) se tiene

$$\prod_{i=1}^{n} [h(t_i)]^{\delta_i} [S(t_i)] \tag{2.17}$$

Esta versión particular de la función de verosimilitud es utilizada cuando la función de densidad de probabilidad tiene una forma complicada. La estimación de parámetros desconocidos en la función de verosimilitud se hace maximizando el logaritmo de esta.

Una aplicación de la ecuación (2.17) se puede hacer a los tiempos de supervivencia de n individuos, $t_1, t_2, \ldots, t_n$, de una distribución exponencial con media λ^{-1}. De esta manera supóngase que tenemos r tiempos de muerte y $n - r$ tiempos de supervivencia censurados a la derecha.

De la distribución exponencial se sabe

$$f(t) = \lambda e^{-\lambda t}$$

$$S(t) = e^{-\lambda t}$$

y sustituyendo en la expresión (2.17), para n observaciones se obtiene

$$L(\lambda) = \prod_{i=1}^{n} \lambda^{\delta_i} e^{-\lambda t_i}$$

y el correspondiente logaritmo de esta función es

$$log[L(\lambda)] = \sum_{i=1}^{n} \delta_i log[\lambda] - \lambda \sum_{i=1}^{n} t_i$$

nótese que $\sum_{i=1}^{n} \delta_i = r$ ya que sólo se tienen r tiempos de muertes, así

$$log[L(\lambda)] = rlog[\lambda] - \lambda \sum_{i=1}^{n} t_i$$

Ahora se necesita identificar el valor $\hat{\lambda}$, para el cual el logaritmo de la función de verosimilitud es un máximo. Pero esto se hace derivando con respecto a λ la última ecuación e igualando a cero, obteniendo lo siguiente

$$\frac{r}{\lambda} - \sum_{i=1}^{n} t_i = 0$$

de modo que $\hat{\lambda}$ es

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^{n} t_i}$$

para el estimador de máxima verosimilitud de λ.

2.7.2. Un modelo para la comparación de dos grupos

Como ya se mencionó anteriormente, es conveniente tener un modelo general para la comparación de dos grupos de tiempos de supervivencia y en este caso se tiene el modelo de riesgos proporcionales. Así supóngase que se tienen dos grupos, el I y el II, y que X es una variable indicadora que toma el valor cero si el individuo esta en el grupo I y uno si el individuo está en el grupo II. Bajo el modelo de riesgos proporcionales, la tasa de muerte al tiempo t del i-ésimo individuo está dada por

$$h_i(t) = e^{\beta x_i} h_0(t) \tag{2.18}$$

en el caso de que $\psi(x_i) = e^{\beta x_i}$ y x_i es el valor de X para el i-ésimo individuo. Como consecuencia de esto, la tasa de muerte al tiempo t para un individuo en el grupo I es $h_0(t)$, y que para un individuo en el grupo II es $\psi h_0(t)$.

Utilizando el método de máxima verosimilitud, supóngase que se tienen n_1 observaciones de individuos en el grupo I que se pueden expresar de la siguiente forma (t_{i1}, δ_{i1}), $i = 1, 2, \ldots, n_1$, donde δ_{i1} toma el valor cero si el tiempo de supervivencia del i-ésimo individuo en este grupo es censurado, y uno si es tiempo de muerte. Similarmente, sea $(t_{i'2}, \delta_{i'2})$, $i' = 1, 2, \ldots, n_2$, las observaciones de los n_2 individuos en el grupo II. Para ejemplificar este

modelo supóngase que cada grupo tiene una distribución exponencial con parámetro λ, así para un individuo en el grupo I la función de densidad y la de supervivencia quedan de la siguiente forma:

$$f(t_{i1}) = \lambda e^{-\lambda t_{i1}}$$

$$S(t_{i1}) = e^{-\lambda t_{i1}}$$

Para los individuos en el grupo II la tasa de muerte es $\psi\lambda$ y sus respectivas funciones de densidad y de supervivencia son:

$$f(t_{i'2}) = \psi\lambda e^{-\psi\lambda t_{i'2}}$$

$$S(t_{i'2}) = e^{-\psi\lambda t_{i'2}}$$

y utilizando la ecuación (2.15) la verosimilitud de $n_1 + n_2$ observaciones, $L(\psi, \lambda)$ es

$$\prod_{i=1}^{n_1} [\lambda]^{\delta_{i1}} e^{-\lambda t_{i1}} \prod_{i'=1}^{n_2} [\psi\lambda]^{\delta_{i'2}} e^{-\psi\lambda t_{i'2}}$$

Si los tiempos de muerte en cada grupo son r_1 y r_2 respectivamente, entonces $\sum_{i=1}^{n_1} \delta_{i1} = r_1$ y $\sum_{i'=1}^{n_2} \delta_{i'2} = r_2$, y el logaritmo de la función de verosimilitud está dado por:

$$log[L(\psi, \lambda)] = r_1 log\lambda - \sum_{i=1}^{n_1} t_{i1} + r_2 log[\psi\lambda] - \psi\lambda \sum_{i'=1}^{n_2} t_{i'2}$$

obteniendo sus respectivas derivadas parciales, se tienen los estimadores de máxima verosimilitud $\hat{\lambda}$ y $\hat{\psi}$ para la comparación de dos grupos.

Capítulo 3

Decremento múltiple

Una metodología para analizar los datos de supervivencia es el *decremento múltiple* (conocido en inglés como *competing risks*), el cual es un modelo relevante cuando hay varios tipos de fracaso, correspondientes a las diferentes causas de muerte, es decir, es un modelo que nos provee de una idea general para la predicción de la ocurrencia de cierto evento en el tiempo t en presencia de otras causas que pueden censurar el evento de interés.

El concepto de decremento múltiple fue introducido por primera vez en el campo de la Oncología. La razón fue por que el tratamiento para el cáncer produce largos tiempos de supervivencia, de esta forma se hizo muy importante no solo tomar en cuenta los efectos del tratamiento sobre el cáncer, sino también la mortalidad por causas no relacionadas, que podrían tener un impacto en el tratamiento. Como un ejemplo hipotético se puede considerar un estudio a hombres mayores de 70 años para los cuales se les diagnosticó cáncer de próstata. Este estudio se realizó por más de 20 años obteniendo que el 66 por ciento de los hombres murieron por causas que no están relacionadas con el estudio, mientras que el 30 por ciento murieron por causas relacionadas con el cáncer de próstata. Actualmente si se tiene un tratamiento agresivo y uno conservador para combatir el cáncer de próstata, los médicos tienen que decidir el tratamiento en base a la esperanza de vida de los pacientes y en los casos de los diagnósticos de cáncer prostático en la etapa temprana, los pacientes son quienes deciden ya que un tratamiento agresivo suele tener efectos secundarios. Pero claro, ésta solo fue una forma de resolver el problema sin utilizar una metodología adecuada, más adelante se mostrará un planteamiento matemático para poder atacar este tipo de problemas.

3.1. Modelo Paramétrico de mezclas para Decremento múltiple

Supóngase que se tiene un conjunto de datos de decremento múltiple que consiste en $t_1, t_2, \ldots, t_n$ tiempos de vida de n individuos y sean $c_1, c_2, \ldots, c_n$ los indicadores de censura, que se muestra a continuación:

$$c_i = \begin{cases} 1, & \text{Si el individuo } i \text{ no está censurado} \\ 0, & \text{Si el individuo } i \text{ es censurado} \end{cases} \tag{3.1}$$

Si c_i=1, i.e,el individuo i muere por alguna causa, se definen los siguientes indicadores:

$$c_{ij} = \begin{cases} 1, & \text{Si el individuo } i \text{ murio por la causa } j \\ 0, & \text{Otro caso} \end{cases} \tag{3.2}$$

para $1 \leq i \leq n$, $1 \leq j \leq J$. También defínase como

$$p_j = P[\text{de que el individuo } i \text{ muera por la causa } j] \tag{3.3}$$

con

$$1 = \sum_{j=1}^{J} p_j \tag{3.4}$$

Asóciese a cada individuo i con las variables aleatorias $B_i, t_{i1}^*, t_{i2}^*, \ldots, t_{iJ}^*$ y u_i, que representan los indicadores de las causas de muerte, los tiempos de vida antes de morir de la causa j y las variables de censura respectivamente del individuo i. También supongase que para cada i, la variable u_i es independiente de $(t_{i1}^*, t_{i2}^*, \ldots, t_{iJ}^*)$ y de B_i, y que son variables aleatorias idénticamente distribuidas. Las u_i son variables aleatorias de censura para los individuos, y que tiene la misma función de distribución en común, G. Ahora se asumirá que las siguientes probabilidades son discretas

$$P[B_i = j] = p_j, 1 \leq j \leq J,$$

donde se puede interpretar como las causas de muerte a los indicadores B_i si el i-ésimo individuo muere por la causa j. De esta manera sea

$$F_j(t) = P[t_{ij}^* \leq t \mid B_i = j] \tag{3.5}$$

que denota la función de distribución condicional de t_{ij}, dada que la muerte es por la causa j. Supóngase que t_i^* son variables aleatorias definidas como

$$t_i^* = t_{ij}^* \; si \; B_i = j \tag{3.6}$$

Cada t_i^* tiene una función de distribución $F(t)$ dada por

$$\begin{aligned} F(t) = P[t_i^* \leq t] &= \sum_{j=1}^{J} P[t_{ij}^* \leq t \mid B_i = j] P[B_i = j] \\ &= \sum_{j=1}^{J} p_j P[t_{ij}^* \leq t \mid B_i = j] \\ &= \sum_{j=1}^{J} p_j F_j(t) \qquad (3.7) \\ &= \sum_{j=1}^{J} p_j (1 - S_j(t)) \\ &= 1 - \sum_{j=1}^{J} p_j S_j(t) \end{aligned}$$

Finalmente las variables aleatorias t_i, c_{ij} y c_i satisfacen que $t_i = min(t_i^*, u_i)$, $c_{ij} = 1_{(t_{ij}^* \leq u_i, B_i = j)}$ y que $c_i = 1_{(t_i^* \leq u_i)} = \sum_j c_{ij}$, para $j = 1, \ldots, J$: $i = 1, 2, \ldots, n$. El t_i^* representa el tiempo de vida verdadero de los individuos, y t_i representa la observación, tal vez censurada de los tiempos de vida.

Las ecuaciones descritas en (3.3) hasta (3.7) provienen de una formulación probabilística para un modelo de mezclas con enfoque en decremento múltiple. Una formulación equivalente puede ser dada utilizando las *funciones de causas específicas*. Este modelo empieza con la tasa de muerte por causa específica

$$\lambda_j(t)dt = P[t_i^* \, \epsilon \, (t, t + dt) \; del \; individuo \; i \; muera \; por \; la \; causa \; j \mid t_i^* > t] \qquad 1 \leq j \leq J \qquad (3.8)$$

donde t_i^* es una variable aleatoria que representa el tiempo de vida del individuo i. Además para $1 \leq j \leq J$ la ecuación (3.7) produce la tasa de muerte asociada con la función de distribución $F(t)$ de t_i^* , así como $\lambda(t) = \sum_j \lambda_j = F'(t)/(1 - F(t))$. De esta manera la probabilidad p_j de que el individuo i muera por la causa j y la función de supervivencia de la distribución $F_j(t)$ bajo la causa j puede ser expresada respectivamente como

$$p_j = \int_0^\infty \lambda_j(t)[1 - F(t)]\, dt \qquad (3.9)$$

$$F_j(t) = \frac{1}{p_j} \int_0^t \lambda_j(t)[1 - F(t)]\, dy \tag{3.10}$$

La función de distribución $p_j F_j(t)$ es referida como la *función de incidencia acumulativa*. De las ecuaciones anteriores se pueden definir las tasas de muerte por causas específicas por

$$\begin{aligned} \lambda_j(t) &= \frac{p_j f_j(t)}{1 - \sum_{j=1}^{J} p_j F_j(t)} \\ &= \frac{p_j f_j(t)}{1 - F(t)}, j = 1, 2, \ldots, J \end{aligned} \tag{3.11}$$

con la misma interpretación que en la ecuación (3.7).

Estimadores de máxima verosimilitud

La verosimilitud para la muestra está dada por

$$\begin{aligned} L_n &= \prod_{i=1}^{n} (p_j f_j(t_i))^{c_i} (P(t_i^* > t))^{1-c_i} \\ &= \prod_{i=1}^{n} (p_j f_j(t_i))^{c_i} \left(1 - \sum_{j=1}^{J} p_j F_j(t_i)\right)^{1-c_i} \end{aligned} \tag{3.12}$$

Dejando t_{ij}=t_i cuando c_{ij}=1, se puede escribir la ecuación anterior como

$$L_n = L_n(\theta) = \prod_{i=1}^{n} \left(\prod_{j=1}^{J} (p_j f_j(t_{ij}))^{c_{ij}} \right) \left(1 - \sum_{j=1}^{J} p_j F_j(t_i)\right)^{1-c_i} \tag{3.13}$$

Los p_j, $1 \leq j \leq J$, constituyen parámetros que se estiman de los datos. Otros parámetros que se especifican en las distribuciones de supervivencia para aquellos que son susceptibles a algún riesgo $\{1, 2, \ldots, J\}$.

3.2. Modelo de Larson y Dinse

Un modelo de mezclas para analizar los datos de decremento múltiple es el de *Larson y Dinse*[8], ya que su formulación está basada en la verosimilitud y que presenta una forma flexible para analizar los datos en presencia de censura. Este modelo supone que la causa de muerte de un individuo es

escogida por un mecanismo estocástico y que el tiempo de supervivencia es una realización de t_{ij}^* de esta forma el modelo se basa en las *funciones de distribución de supervivencia condicionales*, definidas como

$$S_j(t) = P[t_{ij}^* > t \mid B_i = j] \tag{3.14}$$

donde $S_j(t)$ es una función de supervivencia *propia*, i.e, $S_j(t) = 1$ y $S_j(\infty) = 0$.

Bajo esta formulación, las probabilidades de muerte por causa específica definidas en la ecuación (3.5) tenemos que $F_j(t) = 1 - S_j(t)$. De esta forma la función de supervivencia total se puede expresar como

$$S(t) = P[t_{ij}^* > t] = \sum_{j=1}^{j} p_i S_j(t) \tag{3.15}$$

Para permitir la inclusión de efectos de variables explicativas en la función de supervivencia condicional $S_j(t)$ o dicho en otras palabras para permitir su efecto en el componente de la tasa de muerte, Larson y Dinse utilizan el *modelo de riesgos proporcionales de Cox*. Específicamente se supone que la tasa de muerte condicional se caracteriza por la siguiente función de supervivencia:

$$S_j(t) = exp\left[-\Lambda_{0j}(t)e^{\beta_j'\mathbf{x}}\right],\ j = 1, 2, \ldots, J. \tag{3.16}$$

donde $\Lambda_{0j}(t) = \int_0^t \lambda_{0j}(u)\, du$, $\lambda_{0j}(t)$ es una fuerza de mortalidad base, $\mathbf{x}$ es un vector de variables explicativas el cual no tiene a la ordenada, y β_j denota el vector de parámetro por la causa de muerte j. En términos de la fuerza de mortalidad, la suposición de riesgos proporcionales es equivalente a que la fuerza de mortalidad condicional $\lambda_j(t; \mathbf{x})$ toma la siguiente forma

$$\begin{aligned} \lambda_j(t; \mathbf{x}) &= -\frac{d}{dt}[\log S_j(t)] \\ &= \lambda_{0j}(t) \exp[\beta_j'\mathbf{x}],\ j = 1, 2, \ldots, J. \end{aligned} \tag{3.17}$$

La construcción de la función de verosimilitud es muy parecida a la de los modelos de supervivencia univariados. Si $f_j(t)$ denota la función de densidad correspondiente a la j-ésima función de supervivencia condicional, i.e. si $f_j(t) = -dS_j(t)/dt$, entonces un individuo i que fallece por la causa j en el tiempo t_i contribuye con $p_j f_i(t)$ a la función de verosimilitud. Por otra parte un individuo que sobrevive a todas las causas finales del seguimiento y por tanto tiene una observación con censura en el tiempo t_i contribuye con

$S(t_i)$ a la función de verosimilitud. De esta forma la función de verosimilitud correspondiente a n individuos se puede expresar como (e.g.[3],p.147):

$$L_n = L_n(\theta) = \prod_{i=1}^{n} \left(\prod_{j=1}^{J} (p_j f_j(t_{ij}))^{c_{ij}} \right) \left(\sum_{j=1}^{J} p_j S_j(t_i) \right)^{1-c_i} \tag{3.18}$$

Capítulo 4

Manejo de datos faltantes

En la actualidad es muy importante saber cómo manejar una base de datos para una adecuada aplicación en la estadística. Pero al momento en que se está trabajando surgen algunos inconvenientes, como es el caso de que algunos datos no concuerdan como deberían o un problema demasiado grave, el de valores faltantes (en inglés *missing value*). Es en este punto donde entra la imputación que es una forma para manejar valores faltantes.

Como se mencionó anteriormente, el presente trabajo está utilizando un conjunto de datos con valores faltantes de aquí la importancia de trabajar con Imputación. Antes de comenzar con una explicación más detallada de lo que es Imputación se hablará de valores faltantes.

4.1. Datos faltantes

Missing data o *missing values* se refiere a un conjunto de valores faltantes, que son desconocidos en una base de datos. En estadística esta información se obtiene en gran parte a los censos y encuestas. Pero de cualquier modo siempre habrá información incompleta. En los estudios epidemiológicos suelen haber muchos datos faltantes, ya que puede existir la ausencia de participación del paciente (en algunos casos por vergüenza no se contesta lo que se le está pidiendo). La notación en general para los datos faltantes es ? aunque la forma estándar de hacerlo es NA y es como se trabajará en la base de datos del cáncer de próstata de la presente tesis.

4.1.1. Patrones de los datos faltantes

Los datos faltantes pueden presentar diferentes patrones, aunque esto no es un punto importante para la técnica de imputación. Supóngase que

se representan los datos en forma matricial, en donde las filas se refieren a las unidades de observación(en este caso a los pacientes) y las columnas representan las variables que se observaron (en este trabajo se refiere a: la etapa del tumor, grado del tumor, tamaño del tumor, etc.), así los diferentes patrones se muestran en la siguiente figura:

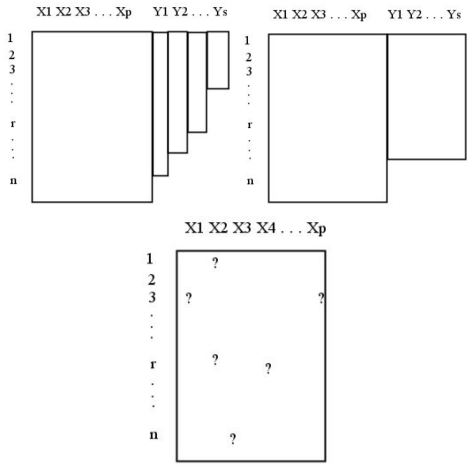

Figura 4.1: Ejemplo de patrones monótonos (Cuando la variable p tiene mayor número de valores observados que la variable $p+1$ y la variable $p+1$ tiene mayor número de valores observados que la variable $p+2$, etc.), multivariados (Cuando las variables $1, 2, \ldots, p$ no tiene datos faltantes y las variables $p+1, p+2, \ldots, p+k$ tienen el mismo número de valores faltantes.) y generales (Cuando los valores faltantes son aleatorios)

Modelos de generación de datos faltantes

Los datos faltantes se pueden clasificar de la siguiente manera:

- **MCAR (Missing Completely at Random)**

Se refiere a que la probabilidad de que una observación sea faltante no depende de los datos observados o no observados. Es decir la ausencia de información es independiente a la matriz de información de datos y a las variables desconocidas de interés.

Un ejemplo de MCAR sería que los datos fueron tomados con un instrumento que no funcione correctamente, por lo tanto puede generar datos perdidos o faltantes.

- **MAR (Missing at Random)**

Decimos que un dato faltante es MAR si la probabilidad de que sea faltante es independiente de los otros valores faltantes pero sí depende de los datos observados.

Un ejemplo de MAR es el siguiente: En las mujeres es menos probable que revelen su peso y/o edad, así la probabilidad de que el dato sea faltante, depende del sexo y no de los otros pesos y/o edades de las demás personas.

- **MNAR(Missing not at Random)**

Decimos que un dato faltante es MNAR cuando no es MAR ni MCAR, Es decir que la probabilidad de que el dato sea faltante depende de los datos faltantes de ellos mismos. Por ejemplo, una persona con sobrepeso no revelará su peso, este dato faltante depende de la variable peso. Este tipo de mecanismo es también llamado no ignorable.

4.2. Formas de tratar a los datos faltantes

Antes de comenzar a introducir el tema de imputación que es un punto clave en esta tesis, se hablará muy brevemente de otras formas para tratar con datos faltantes.

4.2.1. Análisis con datos completos (Listwise)

Este es un método muy sencillo para trabajar, que consiste en eliminar los registros con datos faltantes y así realizar el análisis estadístico solamente con los datos que dispongan de la información completa para todas las variables que se están midiendo. La ventaja de este método es la facilidad en que se puede implementar, pero en realidad conlleva a una gran pérdida de la información y más cuando se tiene una gran cantidad de datos faltantes, y esto puede generar sesgos en las estimaciones de los parámetros.

Al estar trabajando con este método, uno está eliminando información que se asume que poseen los datos completos, y que la pérdida de los datos se generó de manera aleatoria, pero en la realidad muchas veces no es así.

Otra desventaja que se presenta es que este método borra información que puede ser de mucha relevancia por la eliminación de datos faltantes.

4.2.2. Análisis con datos disponibles (Pairwise deletion)

Una forma diferente para trabajar con datos faltantes es el Pairwase deletion ya que este método consiste en utilizar todos los datos que disponga

cada una de nuestras variables, pero la desventaja de este método es que se utilizan distintos tamaños de muestra dependiendo de la variable, claro, se obtienen muy buenos resultados pero cuando se está trabajando con MCAR. Comparado con el método anterior tiene la ventaja de que utiliza toda la información, pero el problema es que los distintos tamaños de la información debilitan la aplicación.

4.2.3. Métodos de imputación

Una de las posibles formas de atacar a los conjuntos de datos con valores perdidos, es utilizar los denominados métodos de imputación.

La palabra **"imputar"** tienes sus orígenes en el latín ***imputo***, que significa atribuir, hacer de cuenta o encargarse, pero en lo que respecta con la estadística, imputación significa "rellenar datos". Imputación es una técnica estadística que consiste en asignar un valor a aquellas variables que tienen datos perdidos utilizando la información contenida en la base de datos, las principales razón por las que se tienen datos faltantes es porque la variable se tomó mal o porque simplemente no se pudo medir. Una razón muy importante por la cual se utiliza este método es porque se obtiene un conjunto de datos completos y consistentes que posteriormente permite hacer un análisis estadístico de acuerdo al objetivo deseado en cada caso.

En el proceso de imputación se bebe escoger cuidadosamente las variables objetivo, las variables auxiliares y los criterios de imputación adecuados. Algunos criterios generales que se pueden considerar son los siguientes:

1. Mantener la distribución de la variable ya que el objetivo de la imputación es que llegue a producir valores muy cercanos a la distribución real.

2. Mantener las correlaciones entre las variables ya que no se tienen que alterar en el proceso de imputación.

3. Consistencia con los otros valores

Los métodos de imputación se pueden clasificar en simples y múltiples. La imputación simple consiste en generar un valor por cada valar faltante con el fin de rellenar esos valores perdidos y así tener una base de datos completa, mientras que la imputación múltiple consiste en asignar "m"valores a cada dato faltante y así generar "m"bases de datos completos, para después estimar los parámetros de interés y posteriormente combinar los resultados obtenidos.

Ventajas y desventajas de la imputación

Una ventaja de utilizar estos métodos de imputación radica en que al rellenar aquellos datos faltantes se reduce el sesgo debido a los valores perdidos, además de que los datos obtenidos son consistentes ya que se generan a partir de los datos observados.

Aunque también suele tener sus desventajas ya que para el análisis posterior no se distingue entre datos originales y datos imputados. Además de que los valores imputados suelen ser buenas estimaciones no podemos asegurar que sean las adecuadas ya que no son reales por lo que no se puede asegurar que mejore con respecto a los datos incompletos porque el proceso de imputación es una forma de simular datos.

Si el método de imputación no es el adecuado lo que puede pasar es que aumente el sesgo y sobre estime la varianza, obteniendo datos inconsistentes y por lo tanto una base de datos no confiable, lo que llevaría a una interpretación falsa de los resultados.

4.2.4. Imputación simple

Como ya se había mencionado en la sección 4.2.3, la imputación simple solo se rellena con un valor V_f posible para cada valor faltante, dicho valor puede ser igual o diferente para cada uno de los valores faltantes.

Un método de imputación simple es el llamado imputación por medias que fue introducido por primera vez en 1932 por Wilks. Este método genera un valor constante, que en este caso es la media de los datos observados para todos los datos faltantes, es decir, para $j = r + 1, \ldots, n$ se tiene:

$$y_j^* = \frac{\sum_{i=1}^{r} y_i}{r}$$

Donde y_i= 1, 2,. . . , n son los datos de los cuales r son observados , $m = n - r$ son faltantes y y_j^* es la media de los datos para todos los valores faltantes. Y la varianza puede ser estimada por

$$\sigma^2 = \frac{(r-1)}{n-1}\sigma_{y_j}^2$$

También para realizar imputación simple podemos encontrar otro método de nombre random regression imputation (RRI), el cual fue propuesto por Buck en 1960. Este propone que la media está condicionada sobre los valores observados y que valores faltantes son remplazados por valores generados

por una regresión de Y sobre los datos observados. De la misma manera se puede encontrar el método de Hot deck para imputación, el cual consiste en una imputación aleatoria donde los valores faltantes son remplazados por los valores observados. Aunque este tipo de imputación conserva la distribución de las variables no es muy conveniente, porque se puede perder la correlación entre las variables.

Un método importante que vale la pena profundizar un poco más y que entra en los diferentes procedimientos de imputación simple es el llamado imputación por regresión que acontinuacion se detallará.Esta forma de imputación se basa en modelos de regresión simple para imputar los datos faltentes de la variable Y por ejemplo. Este método consiste en eliminar las observaciones con datos incompletos y utilizar la regresión para predecir estos valores faltantes. Por ejemplo, supongamos que n es el tamaño de la muestra y considerese a Y como la variable con r datos observados y $n-r$ datos faltantes, además sea $X = (X_1, X_2, \ldots, X_s)$ el conjunto de datos que no presenta valores faltantes, de esta forma si el valor y_i es faltante entonces este valor es imputado mediante un modelo de regresión de la siguiente forma:

$$g\{E[Y]\} = X\beta, Y \sim F$$

Donde g es conocida como una función link y F es una función de distribución.

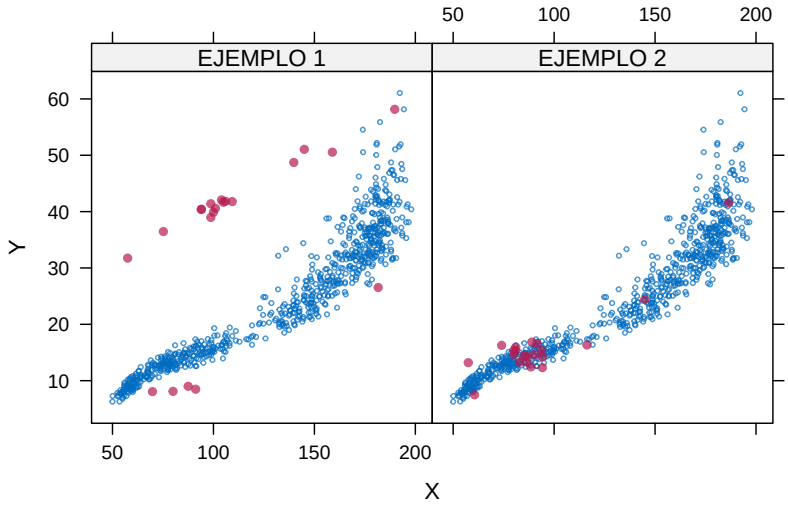

Figura 4.2: Dos maneras distintas de imputar la variable Y, donde los puntos en azul representan los observados y los puntos en rojo los imputados.

En el siguiente sección se detallará sobre el método de imputación múltiple que se ha mencionado anteriormente ya que es parte importante en la presente tesis, por lo que es muy trascendental profundizar en el tema.

4.3. Imputación múltiple

Imputación múltiple es una técnica estadística para analizar bases de datos incompletas, es decir que algunas entradas son valores faltantes. Está consiste en reemplazar cada uno de estos valores faltantes por dos o más valores posibles; esta idea fue propuesta por Rubín [16] en 1977. La Figura 4.3 muestra la idea de este método.

Este método es muy diferente a los ya mencionados anteriormente, ya que genera $m > 1$ valores posibles para cada valor faltante y de esta forma se dispone de m conjuntos de datos completos para después ser analizados y obtener el resultado adecuado. Basicamente imputación multiple requiere de tres pasos:

1. **Imputación.** Generar m valores para cada entrada faltante de una distribución y así generar m bases de datos completos.

2. **Análisis.** Analizar cada conjunto de bases de datos completos, es decir, se harán m análisis.

3. **Resultados.** Integrar los resultados de cada análisis en un resultado final con alguna regla existente para combinar los m resultados.

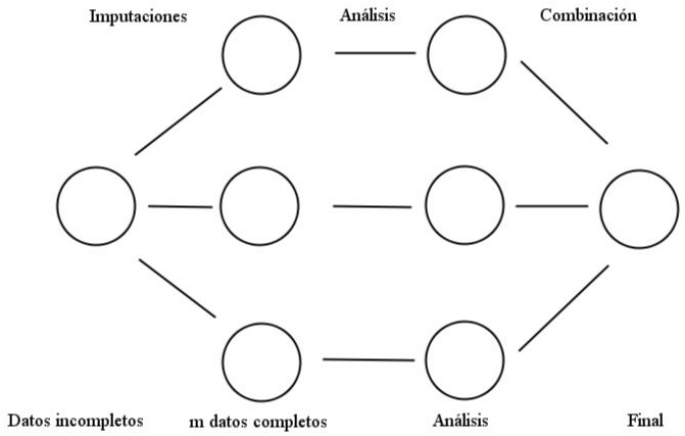

Figura 4.3: Esquema de Imputación múltiple

La idea importante de la imputación múltiple es usar toda la información disponible ya que como se mencionó anteriormente hay métodos que desechan los datos con valores faltantes y al final esa información pudo haber sido muy útil para el análisis posterior de los datos en el problema deseado.

La propuesta original de Rubín no contaba con la formulación para calcular las estimaciones de las combinaciones, pero en 1987 propuso la metodología necesaria para establecer las fórmulas que se obtiene de la combinación de las estimaciones de cada imputación, a la cual se le llamo la regla de Rubín.

La regla de Rubín sugiere que se tome la media de todas ellas, es decir, sean $\widetilde{\theta}_i$ y W_i , con $i = 1, 2, \ldots, m$ las estimaciones deseadas en cada conjunto de datos y las varianzas respectivas a cada estimación para un parámetro θ, de esta manera la estimación combinada es:

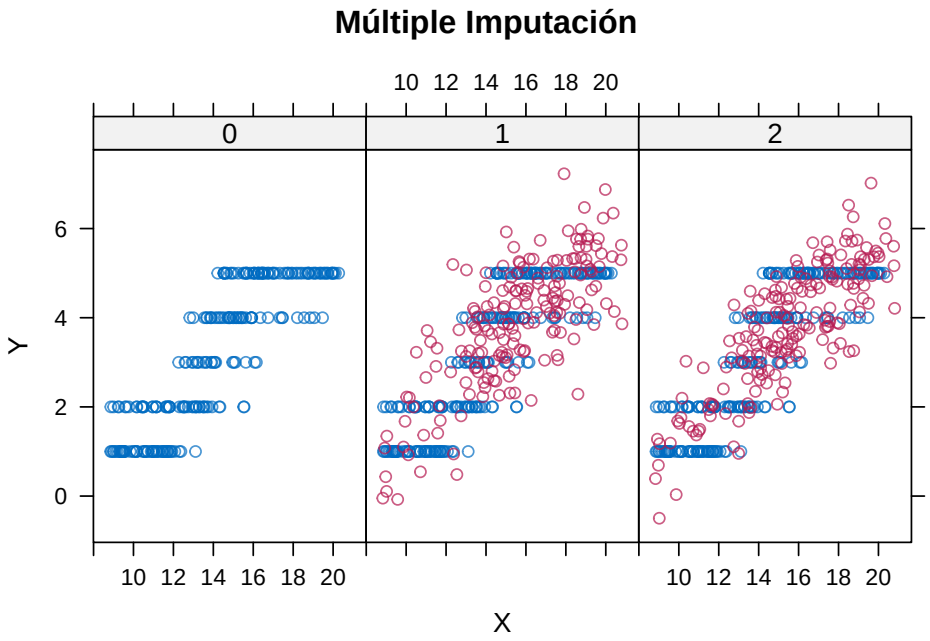

Figura 4.4: Imputación de la variable Y, donde los puntos en azul representan los valores observados y los puntos en rojo los valores imputados

$$\overline{\theta} = \frac{1}{m} \sum_{i=1}^{m} \widehat{\theta}_i$$

La varianza asociada a la estimación anterior tiene dos componentes

- La varianza en cada imputación

$$\overline{W} = \frac{1}{m} \sum_{i=1}^{m} W_i$$

- La varianza entre las imputaciones

$$\overline{B} = \frac{1}{m-1}\sum_{i=1}^{m}(\widehat{\theta_i} - \overline{\theta_m})^2$$

Por lo que la varianza total asociada a la estimación $\overline{\theta_m}$ es

$$T = \overline{W} + \frac{m+1}{m}\overline{B}$$

4.4. Datos incompletos

4.4.1. Causas de datos faltantes

Hay dos causas de datos faltantes: los que son intencionados y los que no lo son. Cuando la causa de datos faltantes es intencionada, la persona encargada de recabar esta información a planeado la falta de estos datos. Por ejemplo, supongamos que un investigador está analizando los tiempos de vida de pacientes con algún tipo de enfermedad mortal, pero decide retirarse antes de que todos sus pacientes mueran por dicha enfermedad, entonces los tiempos vida de los pacientes son censurados y por este motivo son faltantes intencionalmente.

Cuando la causa de datos faltantes no es intencionada, la persona encargada de recabar los datos no ha planeado la falta de estos. Por ejemplo, supongamos que la forma en que se está recabando información es por medio de una encuesta, y que algunas personas se rehusan a contestar una o más preguntas por lo que el encuestador no tiene control sobre los datos faltantes.

4.4.2. Notación

Como anteriormente se ha visto el símbolo "m" es utilizado para denotar el número de las imputaciones. Ahora sea Y una matriz de n x p que contiene el valor de "p" variables y cada una de tamaño n, y sea R una matriz de tamaño n x p la cual contiene 0s y 1s y que se conoce como la *indicadora de respuestas*. Sean y_{ij} y r_{ij}los elementos de Y y R respectivamente, $i = 1, \ldots, n$ y $j = 1, \ldots, p$, entonces:

$$r_{ij} = \begin{cases} 1 & \text{Si } y_{ij} \text{ es observado,} \\ 0 & \text{Si } y_{ij} \text{ si es faltante.} \end{cases} \tag{4.1}$$

Obsérvese que independientemente de que la matriz Y tenga valores faltantes en algunas de sus entradas (inclusive en todas), la matriz R siempre estará completamente observada. Ahora sea Y_{obs} el conjunto de valores en

Y que son observados y sea Y_{mis} el conjunto de valores faltantes en Y, de esta forma $Y = (Y_{obs}, Y_{mis})$. Si $Y = Y_{obs}$ entonces Y no tiene datos faltantes, y si además se conocen los mecanismos en que los datos son faltantes (MAR,MNAR ó MCAR), se podría verificar adecuadamente que las estimaciones de los faltantes son apropiadas, aunque en las aplicaciones interesantes de la realidad eso no sucede.

4.4.3. Mecanismos de datos faltantes

Anteriormente ya se han mencionado los diferentes mecanismos de datos faltantes, pero en esta parte se darán las definiciones en términos matemáticos. De la misma manera consideremos a la matriz R como la indicadora de los datos faltantes de Y, nótese que la distribución de los valores de R depende de $Y = (Y_{obs}, Y_{mis})$ para describir el modelo de los datos faltantes.

La forma general para describir la expresión del modelo de los datos faltantes es

$$P\{R = 0|Y_{obs}, Y_{mis}\} \tag{4.2}$$

De acuerdo a lo anterior, si el mecanismo es **MCAR** entonces

$$P\{R = 0|Y_{obs}, Y_{mis}\} = P\{R = 0\} \tag{4.3}$$

es decir, que no depende de los observados y ni de los faltantes. Si el mecanismo es **MAR** entonces

$$P\{R = 0|Y_{obs}, Y_{mis}\} = P\{R = 0|Y_{obs}\} \tag{4.4}$$

es decir, que la probabilidad de que los datos sean faltantes depende de la información de los datos observados. Por último si el mecanismo es **MNAR** entonces

$$P\{R = 0|Y_{obs}, Y_{mis}\} \tag{4.5}$$

es decir, no simplifica la expresión (4.2), ya que la probabilidad de ser perdido depende de los datos observados y de los faltantes en si mismos.

Ejemplo. Supongamos que se quieren simular los diferentes mecanismos de datos faltantes. Sea $Y = (Y_a, Y_b)$ el conjunto de datos que fueron creados por alguna distribución. Entonces si se quieren crear datos faltantes en Y_b suponiéndose que se está utilizando el siguiente modelo

$$P\{R_b = 0\} = \tau_0 + \tau_1 \frac{exp(Y_a)}{1 + exp(Y_a)} + \tau_2 \frac{exp(Y_b)}{1 + exp(Y_b)} \tag{4.6}$$

Sea $\tau = (\tau_0, \tau_1, \tau_2)$,entonces para **MCAR** se utilizará $\tau = (r, 0, 0)$ con $0 \leq r \leq 1$, para **MAR** se toma $\tau = (0, 1, 0)$ y para **MNAR** se toma $\tau = (0, 0, 1)$, y posteriormente usar un modelo bernoullí para determinar si se tomará como faltante u observada.

4.4.4. Ignorabilidad

En la sección 4.4.3 se hablo de un ejemplo para generar mecanismos de datos faltantes, el cual dependía de los parámetros de τ, aunque en las aplicaciones en general estos valores son desconocidos, y de alguna manera el interés primordial es poder continuar sin la necesidad de conocer estos valores.

Para resolver esta problemática Little y Rubin [8] utilizaron la verosimilitud de θ basado en los valores observados Y_{obs} en el caso de que se ignoren los mecanismos, la cual es proporcional a la densidad de probabilidad marginal de Y_{obs}, es decir:

$$L(\theta|Y_{obs}) \propto f(Y_{obs}|\theta) \tag{4.7}$$

donde $f(Y_{obs}|\theta) = \int f(Y_{obs}, Y_{mis}|\theta)\, dY_{mis}$ y $f(Y_{obs}, Y_{mis}|\theta)$ denota la densidad de distribución conjunta de Y_{obs} y Y_{mis}.

Ahora en el caso de la imputación las variables observadas constan de (Y_{obs}, R), así su distribución está dada por

$$f(Y_{obs}, R|\theta, \tau) = \int f(Y_{obs}, Y_{mis}|\theta) f(R|Y_{obs}, Y_{mis}, \tau)\, dY_{mis} \tag{4.8}$$

con $f(R|Y_{obs}, Y_{mis}, \tau)$ la distribución conjunta del modelo de los datos faltantes y la función de verosimilitud está dada por:

$$L(\theta|Y_{obs}, R) \propto f(Y_{obs}, R|\theta) \tag{4.9}$$

La cuestión ahora es como hacer inferencias acerca de θ basadas en la verosimilitud (4.9), y de qué forma se puede omitir los mecanismos de datos faltantes para hacer inferencias acerca de (4.7), Notése que si se supone MAR en (4.8) se tiene:

$$f(Y_{obs}, R|\theta) = f(Y_{obs}|\theta) f(R|Y_{obs}) \tag{4.10}$$

y que bajo esta suposición las inferencias de θ en $L(\theta|Y_{obs}, R)$ serian las mismas que $L(\theta|Y_{obs})$ al cual se denomina mecanismo ignorable.

Ejemplo. Supóngase que se tiene una variable $Y = (Y_{obs}, Y_{mis})$ con $Y_{obs} = (y_1, \ldots, y_m)$ y $Y_{mis} = (y_{m+1}, \ldots, Y_n)$. Para simplificar los cálculos también supóngase que las y_i son variables aleatorias exponenciales, así la función de distribución conjunta es:

$$f(Y|\theta) = (\theta)^{-n} exp\left(-\sum \frac{y_i}{\theta}\right) \tag{4.11}$$

La verosimilitud, ignorando los mecanismos de datos faltantes, es proporcional a la distribución de Y_{obs} dado θ y está dada por:

$$f(Y_{obs}|\theta) = (\theta)^{-m} exp\left(-\sum_1^m \frac{y_i}{\theta}\right) \tag{4.12}$$

Ahora sea $R = (R_1, \ldots, R_n)$, $R_i = 1$, $i = 1, \ldots, m$ y $R_i = 0$, $i = m+1, \ldots, n$. Supóngase que cada valor observado tiene probabilidad ψ, así

$$f(R|Y_{obs}) = \psi^m$$

y

$$f(Y_{obs}, R|\theta) = \psi^m(\theta)^{-m} exp\left(-\sum_1^m \frac{y_i}{\theta}\right)$$

Si ψ y θ son distintos, las inferencias de θ pueden ser basadas en $f(Y_{obs}|\theta)$, ignorando el mecanismo de datos faltantes. En particular la estimación de θ es simplemente $\sum_1^m \frac{y_i}{m}$, la media de los valores de Y.

Ahora la importancia de esta sección radica en la necesitan dibujar observaciones sintéticas de la distribución de los datos faltantes, dados los datos observados y un mecanismo de generar datos faltantes para las imputaciones, dicha distribución es denotada por:

$$P\{Y_{mis}|Y_{obs}, R\} \tag{4.13}$$

Pero lo que se quiere es un modelo que no necesite explícitamente el mecanismo de datos faltantes, es decir,

$$P\{Y_{mis}|Y_{obs}, R\} = P\{Y_{mis}|Y_{obs}\} \tag{4.14}$$

Rubín [17] demostró que se si se tiene un mecanismo ignorable entonces se cumple (4.14), lo que implica

$$P\{Y|Y_{obs}, R = 1\} = P\{Y|Y_{obs}, R = 0\}, \tag{4.15}$$

así la distribución de los datos Y es la misma para los no observados y los observados. De esta manera se puede utilizar un modelo en base a los observados para imputar los valores faltantes que básicamente es lo que implica el mecanismo de **MAR**.

4.4.5. Elección del número de imputaciones

Múltiple imputación es una técnica atractiva porque da inferencias validas para valores pequeños de m. Algunos investigadores en la actualidad toman entre tres y diez imputaciones porque son suficientes para obtener resultados aceptables, ya que para definir el número de imputaciones a usar Rubín [17] utilizó la eficiencia de m imputaciones respecto a un número infinito de imputaciones, es decir, la eficiencia relativa, que es aproximadamente:

$$ER = \left(1 + \frac{\gamma}{m}\right)^{-1} \tag{4.16}$$

con

$$\gamma = \frac{r + \frac{2}{d+3}}{r+1}$$
$$d = (m-1)\left(1 + \frac{m\overline{W}}{(m+1)\overline{B}}\right)^2$$
$$r = \frac{(1+\frac{1}{m})\overline{B}}{\overline{W}}$$

donde γ es conocida como la fracción de información faltante. De esta manera la eficiencia relativa para diferentes valores de γ es mostrada en el cuadro (4.1).

Cuadro 4.1: Eficiencia para las diferentes imputaciones y para los diferentes valores de la fracción de información faltante γ

Imputaciones	γ=0.1	γ=0.3	γ=0.5	γ=0.7	γ=0.9	γ=0.99
1	90.91	76.92	66.67	58.82	52.63	50.25
2	95.24	86.96	80.00	74.07	68.97	66.89
3	96.77	90.91	85.71	81.08	76.92	75.19
5	98.04	94.34	90.91	87.72	84.75	83.47
7	98.59	95.89	93.33	90.91	88.61	87.61
9	98.90	96.77	94.74	92.78	90.91	90.09
10	99.01	97.09	95.24	93.46	91.74	90.99
25	99.60	98.81	98.04	97.28	96.53	96.19
50	99.80	99.40	99.01	98.62	98.23	98.06
100	99.90	99.70	99.50	99.30	99.11	99.02

En base al cuadro (4.1) se puede notar que si se quiere tener una eficiencia relativamente razonable hay que tomar $m \geq 10$, y para el presente trabajo se tomara $m = 10$.

4.5. Ventajas y desventajas

Las ventajas que presenta el método de imputación multiple con respecto a la imputación simple son los siguientes:

- La imputación múltiple mejora la eficiencia de los estimadores ya que minimiza los errores estándar.
- Obtiene inferencias razonables mediante la combinación de las inferencias obtenidas en las bases de datos completas.
- Permite usar directamente las inferencias de las bases de datos completas generadas por las m imputaciones.

Por otro lado las desventajas que se pueden observar es que son procesos que requieren un mayor esfuerzo, mayor tiempo para el análisis, y mayor trabajo computacionalmente, aunque claro, estas desventajas no suelen ser tan malas en particular para el análisis cuando m es pequeña.

4.6. Método de Imputación Múltiple

De la misma manera que antes, supóngase que se tiene un conjunto de datos de un estudio que es representado por una matriz Y de tamaño n x p la cual tiene datos faltantes Y_j con $j = 1, \ldots, p$ las variables de Y y donde n representa el número de individuos en el estudio .Supóngase que las variables Y_{mK} con $K = 1, \ldots, n_1$ tiene datos faltantes con $n_1 \leq p$, así el número de variables que están completamente observadas son Y_{oL} con $L = 1, \ldots, n_2$ y $n_2 = p - n_1$, y sea $Y_{-j} = (Y_1, Y_2, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p)$ el complemento de la variable Y_j.

El proceso de imputación múltiple comienza con la elección de un modelo de imputación $P(Y_{mK}|Y_{-mK}, R)$ para cada variable con datos faltantes, para después rellenar cada dato faltante Y_{mK} con un valor aleatorio que fue observado de ésta.

Para la primer variable con datos faltantes, digamos Y_{m1}, se utiliza un modelo de regresión lineal sobre todas las otras variables(en este caso la regresión se hará con las variables Y_{-m1}), restringido a todos los individuos para los cuales Y_{m1} es observada. Los valores faltantes de la variable Y_{m1} son remplazados por simulaciones correspondientes al modelo de imputación. Entonces para la siguiente variable con datos faltantes, digamos Y_{m2}, se utilizará un modelo de regresión sobre todas las otras variables Y_{-m2}, restringido a los individuos para los cuales Y_{m2} es observada y usando los valores imputados de Y_{m1}. De la misma manera los valores de Y_{m2} son remplazados por las simulaciones correspondientes al modelo de imputación de Y_{m2}. El proceso se repite para todas las demás variables con datos faltantes Y_{mK}, este proceso es conocido como un ciclo. Van Buuren y Brand [40] proponen usar entre 5 y 10 ciclos por imputación ya que este último simuló diferentes mecanismos de

datos faltantes en diferentes tipos de variables para después utilizar el método de imputación múltiple, obteniendo estimaciones satisfactorias de los parámetros de interés.Para finalizar el procedimiento se repite m veces para generar las m bases de datos rellenas con las imputaciones.

Una característica de la imputación múltiple es la habilidad para manejar diferentes tipos de variables (continuas, binarias, categóricas, etc.) porque cada variable es imputada usando su propio modelo de imputación. En la siguiente parte se detallarán los diferentes modelos de imputación de acuerdo al tipo de variable.

4.7. Imputación para diferentes tipos de variables

En esta parte se introducirá brevemente los diferentes enfoques para imputar datos faltantes en variables continuas, binarias y categóricas, aunque para el presente trabajo solo se estará utilizando imputación para variables categóricas.

En el proceso de imputación múltiple se mencionó la elección de un modelo de imputación de acuerdo al tipo de variable, los más comunes son los siguientes:

- **Continua** Para una variable continua Y se trabaja con un modelo de regresión normal lineal, $Y = U\beta + e$, donde U es conocida como la matriz de diseño de los datos cuyas columnas son las variables que se miden en el estudio y los renglones representan a los individuos, el vector e tiene una distribución normal con media cero y varianza $\sigma^2 I$ con I como la matriz identidad.

- **Binaria** Cuando Y es una variable binaria, lo que se utiliza es un modelo de regresión lineal Logístico de Y sobre U, $logic[P[Y = 1|U]] = U\beta$

- **Categóricas** Es una variable que puede tomar k distintos valores. $j = 1, 2, \ldots, k$, sea $\pi_j = P[Y = j|U]$ donde este modelo utiliza una regresión de la siguiente forma $log(\pi_j/\pi_r) = U\beta_j$ para $j \neq r$ y r fijo.

En el problema del cáncer de próstata sólo se imputarán variables categóricas, por lo que se detallará la forma en que se crean las imputaciones. Para simplificar los cálculos supóngase que $r = k$, de esta manera para estimar los coeficientes β_j se utilizará

$$log(\pi_j/\pi_k) = U\beta_j$$

con $\pi_j = \dfrac{exp[U\beta_j]}{1+\sum_{j\neq k} exp[U\beta_j]}$, $\pi_k = \dfrac{1}{1+\sum_{j\neq k} exp[U\beta_j]}$

Así sea V la matriz de covarianzas de $\beta = (\beta_1, \beta_2, \ldots, \beta_{k-1})$ y T su descomposición de Cholesky, de modo que la creación de las imputaciones sean de la siguiente manera:

1. Defínase $\beta_{j^*} = \beta_j + Tz$, con z un vector generado de una distribución normal estándar.

2. Sea U_{mis} la notación de los renglones de U donde Y es faltante y sea $P_{j^*} = exp[U_{mis}\beta_{j^*}]/(1+\sum_j exp[U_{mis}\beta_{j^*}])$, donde $j = 1, 2, \ldots, k-1$ y $P_k = 1 - \sum_j P_{j^*}$.

3. Sean $R_0 = 0$, $R_j = \sum_j^{k-1} P_{j^*}$ y $R_k = 1$ las sumas acumulativas de las probabilidades. Para imputar los valores se genera un número u uniforme y se toma el j-ésimo valor de Y si $R_{j-1} \leq u < R_j$

Capítulo 5

Aplicación

En este capítulo se utilizara un análisis inicial de los datos del cáncer de próstata utilizando la función de incidencia acumulativa basados en el status, el tiempo y en los algunos casos en el tipo de tratamiento en el que fueron sometidos los pacientes, para después proponer un procedimiento estadístico para estimar las probabilidades de fallecer dentro de un periodo arbitrario de tiempo de acuerdo al tipo de tratamiento y en presencia de variables auxiliares, las cuales deben de incluir el estado de salud del paciente. Los parámetros del modelo podrán ser empleados para construir criterios que ayuden a decidir qué tipo de tratamiento es el más adecuado para incrementar las probabilidades de supervivencia de cada paciente.

La dificultad de este estudio es que algunas variables tienen datos faltantes (missing values), y de aquí la importancia de trabajar con un método estadístico para el manejo de base de datos con valores faltantes como lo es la Imputación Múltiple del que ya se hablo.Dado que el estudio consta de datos de análisis de supervivencia el modelo propuesto es el de Larson y Dinse que se basa en el Decremento Múltiple el cual es muy relevante cuando hay diferentes tipos de fracaso para diferentes causas de muerte que pueden censurar el evento de interés y que presenta una forma flexible para analizar la presencia de censura además de que su formulación utiliza el modelo de riesgos proporcionales de Cox.

El proceso de la aplicación será el siguiente:

- **Proceso de imputacion**
- **Análisis**
- **Resultados**

En la parte de la imputación se describirá algunos puntos importantes que se tomaron en cuenta para este método y los resultados de las variables impu-

tadas (etapa y grado). Después se procederá con el Análisis de las imputaciones utilizando el modelo propuesto para finalizar con la interpretación de los resultados.

5.0.1. Análisis inicial

Antes de comenzar con el modelo propuesto, se hará un análisis inicial sobre las probabilidades de que los pacientes mueran por la causa $j(j = 1, 2)$ antes de un cierto tiempo t_0 y para ello se utilizaran las funciones de incidencia acumulativas que a continuación se detallara.

Función de Incidencia acumulada

La función de incidencia acumulada se puede definir como la probabilidad de que pase un suceso en un cierto periodo, en este caso el suceso se refiere al fracasos por la causa $k(k = 1, 2)$ y es estimada de la siguiente forma:

$$\widetilde{I_k(t)} = \sum_{j|t_j<t} \widehat{S(t_{j-1})} \frac{d_{kj}}{n_j}$$

Donde d_{kj} es el número de fracasos por k al tiempo t_j, n_j es el número total de casos que están en riesgo al tiempo t_j y $\widehat{S(t_{j-1})}$ es el estimador de Kaplan-Meir el cual aproxima a la función de supervivencia al tiempo t_{j-1} y se calcula de la siguiente manera:

$$\widehat{S(t_{j-1})} = \prod_{j|t_j<t} \frac{n_j - d_{kj}}{n_j}$$

De acuerdo a la base de datos que se tiene, se puede estimar la probabilidad de que un paciente muera por cáncer de próstata o por otras causas a un determinado tiempo como se muestra en la Figura 5.1

En la Figura 5.1 se puede dar cuenta que una vez que se haya detectado a un paciente con cáncer de próstata es más probable que este muera por otra causa ajena a la enfermedad. En la Figura 5.2 se puede notar la probabilidad de que un paciente muera por cáncer de próstata dado que está tomando uno de los tratamientos.

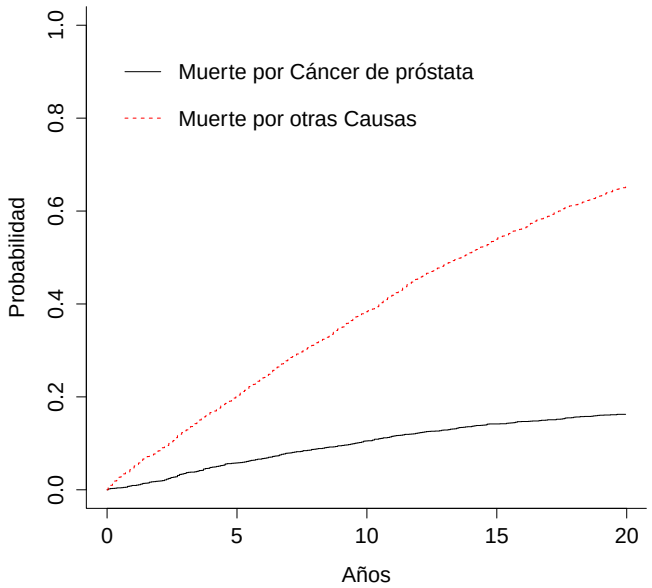

Figura 5.1: Estimación de la función de incidencia acumulativa para la muerte por cáncer de próstata y la muerte por otras causas.

Observese que una persona que no se somete a el tratamiento I (no quirúrgico) tiene una mayor probabilidad que muera por Cáncer de próstata desde el momento que se le detecto la enfermedad mientras que es más probable que una persona muera por otras causas ajenas al cáncer si se le práctica una cirugía simple como se ve en la Figura 5.3

El análisis de esta sección se basa en el status, en el tiempo y en las Figuras 5.2, 5.3 en el tipo de tratamiento en el que fueron sometidos los pacientes del conjunto de datos en general, pero más adelante se hablara sobre la técnica utilizada para manejar datos incompletos, que posteriormente serán utilizados para trabajar con un modelo estadístico que permita incluir las demás variables en este estudio ya que de alguna forma influyen en la elección de un tratamiento adecuado para poder obtener la probabilidad de sobrevivir de cáncer de próstata a un tiempo específico de cada paciente y no en general como lo muestran las gráficas de esta sección.

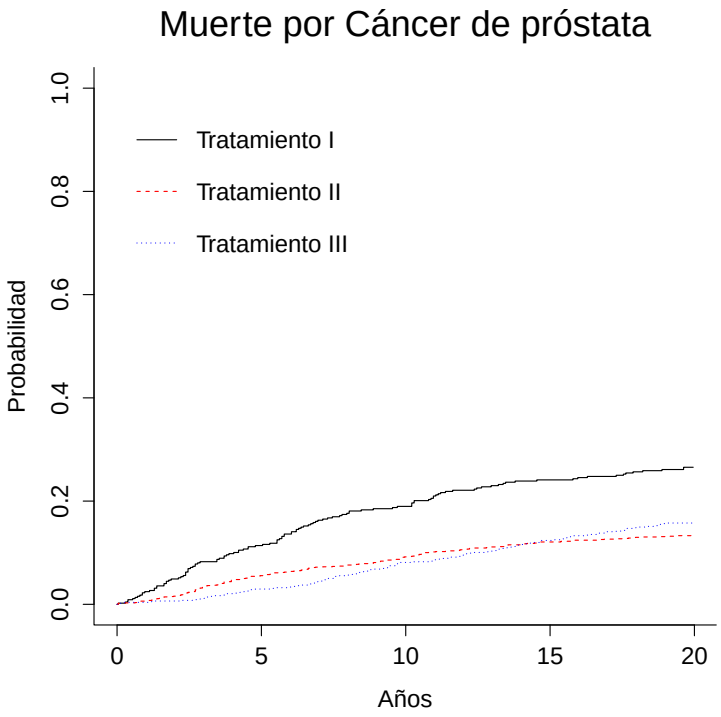

Figura 5.2: Estimación de la función de incidencia acumulativa de la muerte por cáncer de próstata para los distintos tratamiento.

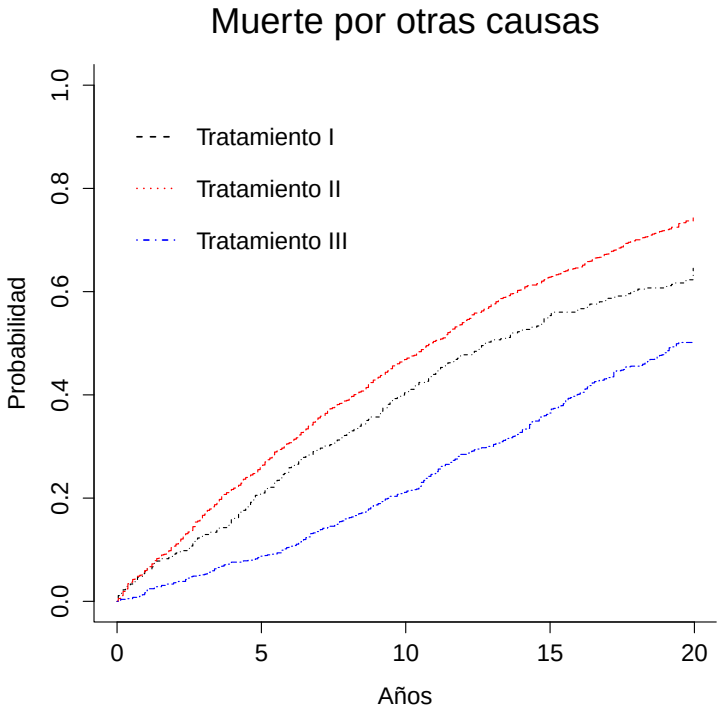

Figura 5.3: Estimación de la función de incidencia acumulativa para la muerte por causas ajenas al cáncer de próstata por tratamiento.

5.1. Proceso de Imputación

Anteriormente ya se había dado una descripción de las diferentes variables que se tiene en la base de datos, de las cuales la variable etapa del tumor y la variable grado del tumor tienen datos faltantes. La etapa del tumor tiene 643 datos faltantes mientras que la variable gardo del tumor tiene 160 datos faltantes, de los cuales 116 datos son faltantes en ambas variables como se muestra en la Figura 5.4.

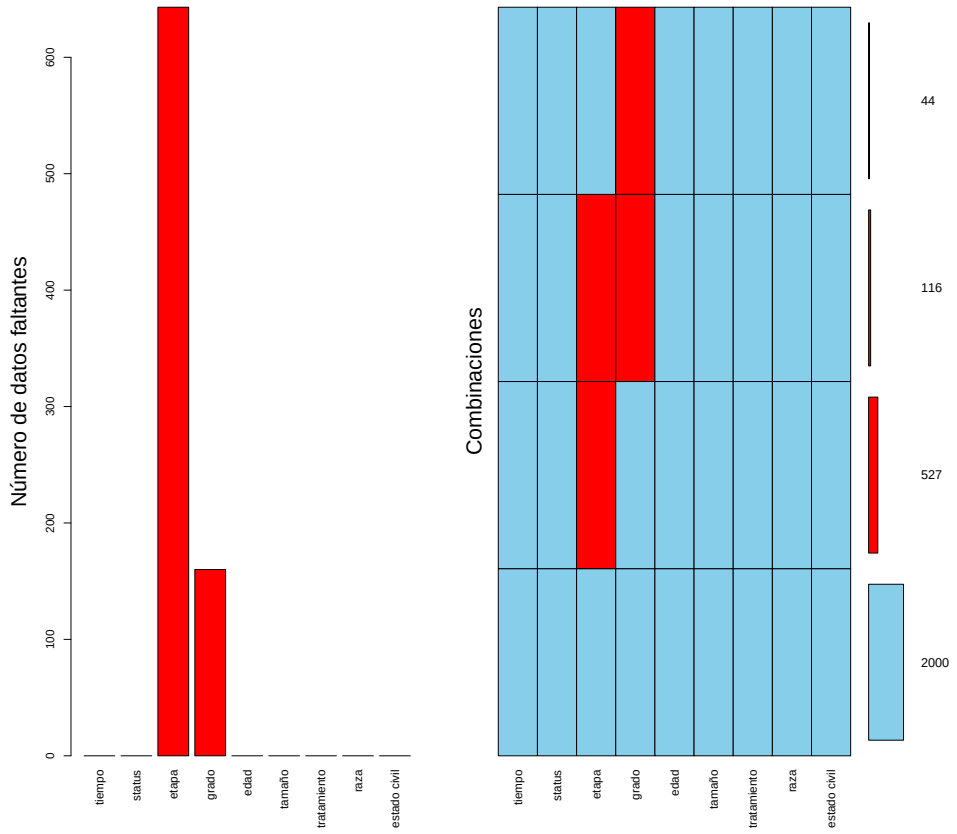

Figura 5.4: En el grafico de la izquierda se puede observar el número de datos faltantes por variable mientras que en el grafico de la derecha se puede observar el número de combinaciones de datos faltantes con las demás variables.

Otra forma de visualizar los datos es directamente de la matriz de información como se ilustra en la Figura 5.5. Básicamente la idea de esta figura es reescalar los valores de las variables continuas en un intervalo del 0 al 1 para representarlos en escala de grises, donde colores obscuros representan valores altos de la variable y los colores claros representan valores bajos de la variable. En el caso de las variables categóricas se asigna un color en la

escala de grises para representar los diferentes valores. Si una variable tiene datos faltantes, entonces estos serán representados en color rojo.

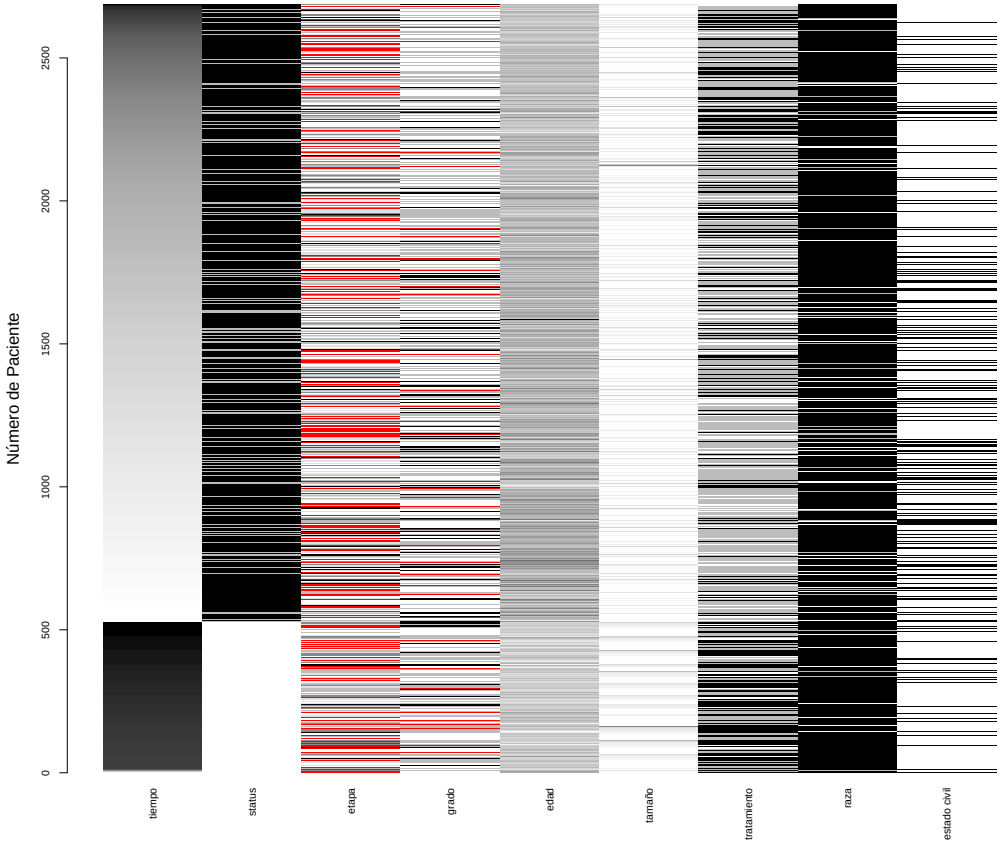

Figura 5.5: Visualización de la base de datos del Cáncer de Próstata

Una parte fundamental en proceso de imputación es ver qué variables se utilizaran como predictores para la imputación de los datos faltantes ya que variables completamente observadas son mejores predictores que las variables con datos faltantes. Para resolver lo anterior se utilizan las frecuencias por pares de variables (Y_j, Y_k) con datos faltantes de la siguiente forma:

1. El número de pares (Y_j, Y_k) para los cuales Y_j y Y_k son observados (a esta forma se le denomina frecuencia rr)

2. El número de pares (Y_j, Y_k) para los cuales Y_j es observada y Y_k es faltante (frecuencia rm)

3. El número de pares (Y_j, Y_k) para los cuales Y_j es faltante y Y_k es observada (frecuencia mr)

4. El número de pares (Y_j, Y_k) para los cuales Y_j y Y_k son faltantes (frecuencia mm)

En el caso de los datos utilizados en esta tesis se tiene que las diferentes frecuencias son: En el caso de los datos utilizados para el cáncer de próstata las frecuencias se muestran el cuadro 5.1

Cuadro 5.1: Frecuencias de los datos del cáncer de próstata

frecuencia rr									
	tiempo	status	etapa	grado	edad	tamaño	tratamiento	raza	civil
etapa	2044	2044	2044	2000	2044	2044	2044	2044	2044
grado	2527	2527	2000	2527	2527	2527	2527	2527	2527
frecuencia rm									
	tiempo	status	etapa	grado	edad	tamaño	tratamiento	raza	civil
etapa	0	0	0	44	0	0	0	0	0
grado	0	0	527	0	0	0	0	0	0
frecuencia mr									
	tiempo	status	etapa	grado	edad	tamaño	tratamiento	raza	civil
etapa	643	643	0	527	643	643	643	643	643
grado	160	160	44	0	160	160	160	160	160
frecuencia mm									
	tiempo	status	etapa	grado	edad	tamaño	tratamiento	raza	civil
etapa	0	0	643	116	0	0	0	0	0
grado	0	0	116	160	0	0	0	0	0

La importancia de las frecuencias anteriores radica en I_{jk} la proporción de casos que se pueden utilizar para imputar la variable Y_j de la variables Y_k definida por Van Buuren [10] como:

$$I_{jk} = \frac{\sum_i^n (1 - r_{ij}) r_{ik}}{\sum_i^n (1 - r_{ij})} \tag{5.1}$$

Esta cantidad también es conocida como *inbound* y es interpretada como el número de pares (Y_j, Y_k) con Y_j faltante y Y_k observada, dividida por el número total de casos faltantes en Y_j. Si la variable Y_k es observada para todos los valores donde Y_j es faltante, entonces la proporción de casos que se pueden utilizar I_{jk} es igual a 1. El *inbound* es utilizado como un predictor Y_k de la variable Y_j, donde valores altos de I_{jk} son preferibles. El *inbound* también puede calcularse como $\frac{mr}{mr+mm}$ y en el caso de los datos de Cáncer de próstata se muestran en el cuadro 5.2.

Cuadro 5.2: Inbound para los datos del Cáncer de próstata

	tiempo	status	etapa	grado	edad	tamaño	tratamiento	raza	civil
etapa	1	1	0	0.82	1	1	1	1	1
grado	1	1	0.28	0	1	1	1	1	1

De la tabla anterior se puede observar que la variable *etapa* no es buen predictor de la variable *grado* ya que su valor es 0,28, en cambio la variable *grado* si es un predictor de la variable *etapa* ya que su valor es de 0,82 y las demás variables son muy buenas predictoras ya que no tienen datos faltantes. El I_{jk} mide qué tan bien las entradas con datos faltantes de la variable Y_j están conectadas con la variable Y_k. El I_{jk} mide qué tan bien las entradas con datos faltantes de la variable Y_j están conectadas con la variable Y_k.

Otros valores son el influx y el outflux definidos por Van Buuren [23], los cuales describen que tan bien cada variables está conectada con las otras.

El influx I_j está definido como:

$$I_j = \frac{\sum_j^p \sum_k^p \sum_i^n (1 - r_{ij}) r_{ik}}{\sum_k^p \sum_i^n (1 - r_{ij})}, \tag{5.2}$$

y es igual al número de pares (Y_j, Y_k) con Y_j faltantes y Y_k observada, dividido entre el número total de pares. Obsérvese que el valor de I_j depende de la proporción de datos faltantes en la variable Y_j, ya que si Y_j es completamente observada entonces $I_j = 0$, en cambio si es completamente faltante $I_j = 1$. Para dos variables con la misma proporción de datos faltantes, la variable con el $influx$ más grande esta mejor conectada con los datos observados y será más fácil de imputar. De manera similarel $outflux$ O_j se define como:

$$O_j = \frac{\sum_j^p \sum_k^p \sum_i^n (1 - r_{ik}) r_{ij}}{\sum_k^p \sum_i^n (1 - r_{ij})} \tag{5.3}$$

La cantidad O_j es el número de pares (Y_j, Y_k), con Y_j observada y Y_k faltante, dividido entre el número total de datos faltantes. El $outflux$ es un indicador de que la variable Y_j se puede utilizar para imputar otras variables. El $outflux$ de una variable completamente observada es igual 1, sin embargo para una variable completamente faltante es igual a 0. Obsérvese que para dos variables con la misma proporción de datos faltantes, la variable con $outflux$ más alto está mejor conectada con las demás, y así es mejor predictor para imputar el resto de las variables.

En la práctica si se grafica el $outflux$ versus el $influx$ para cada variable como se muestra en la figura 5.6, se tiene que $I_j + O_j \leq 1$ y además las variables que se encuentran cercanas a la diagonal están mejor conectadas a

los otros datos. Las variables que se encuentren cerca del origen en la gráfica y que no sean del interés en la investigación se deben de remover para el proceso de imputación.

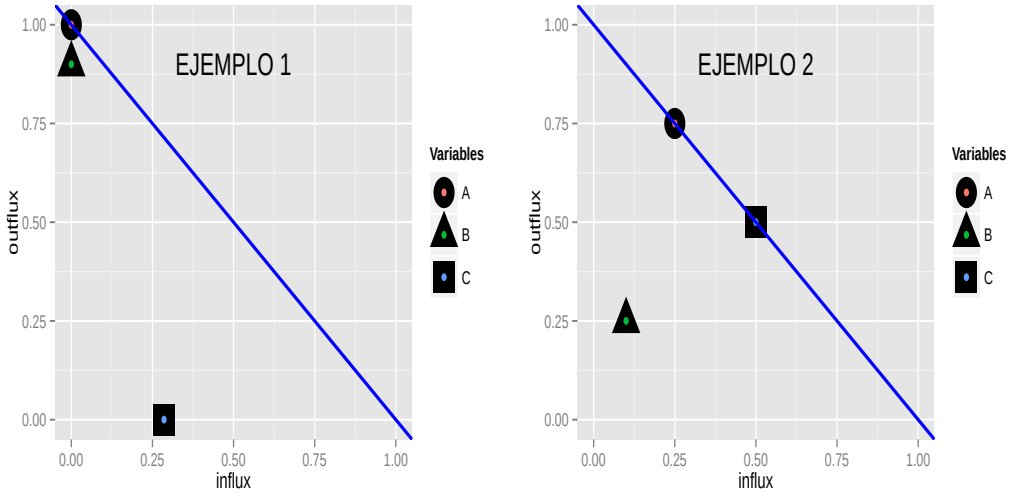

Figura 5.6: Influx versus outflux correspondiente a dos diferentes bases de datos.

En el caso de lo datos del cáncer de próstata se calculó el influx y el outflux, Cuadro 5.3

Cuadro 5.3: influx y outflux para los datos del cáncer de próstata

	influx	outflux
tiempo	0	1
status	0	1
etapa	0.22	0.05
grado	0.05	0.66
eda	0	1
tamaño	0	1
tratamiento	0	1
raza	0	1
civil	0	1

Se puede observar que la variable *etapa* tiene un valor de $outflux$ bajo como se muestra en la Figura 5.7, y esto significa que esta variable no es buen predictor de las demás variables con datos faltantes (en este caso sólo para la variable *grado*), además también se había mencionado que el valor del *inbound* del *grado* con la *etapa* era pequeño por lo que no es buen predictor de esta variable, por lo tanto para el proceso de imputación no se utilizará la variable *etapa* como predictor para imputar a la variable *grado* pero para

imputar a la variable *etapa* si se utilizará la variable *grado* ya que los valores del outflux e inbound son altos.

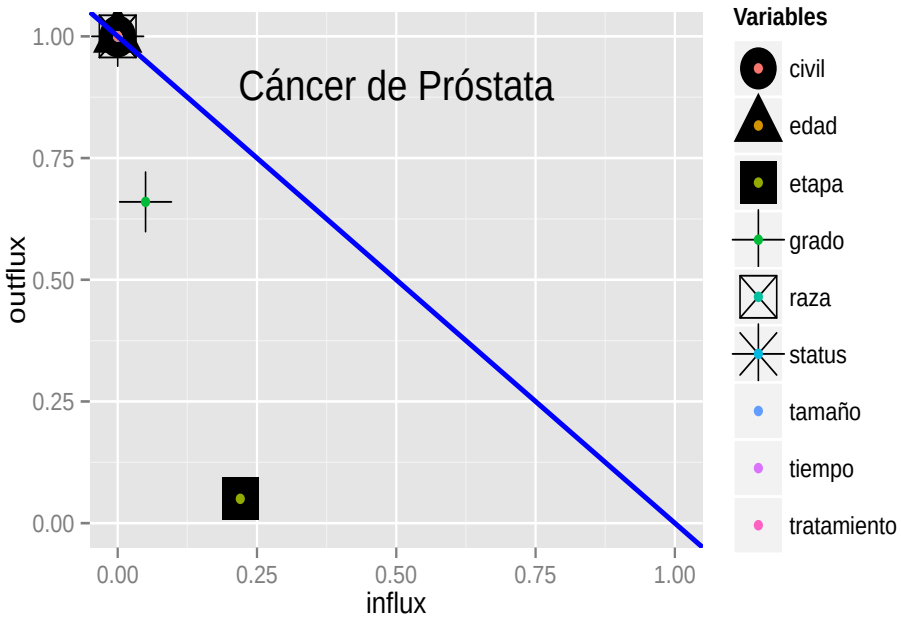

Figura 5.7: Influx versus Outflux de los datos de Cáncer de Próstata

De la discusión anterior podemos concluir que las variables que se utilizarán para el proceso de imputación de la variable *grado* son *tiempo*, *status*, *edad*, *tamao*, *tratamiento*, *raza* y *civil*; mientras que para la variable *etapa* se utilizarán las variables: *tiempo*, *status*, *edad*, *tamao*, *tratamiento*, *raza* ,*civil* y *grado*.

Para generar las imputaciones se utilizará la librería **mice** del paquete estadístico **R**, que dibuja las imputaciones utilizando el proceso descrito en el capítulo anterior, indicándole el número de imputaciones m que se desean, los predictores de cada variable con datos faltantes y automáticamente elige un modelo de imputación de acuerdo al tipo de variable (en el caso del presente trabajo solo se trabajará con el modelo para variables categóricas, también conocido como modelo multinomial logístico), aunque esto se puede cambiar si el usuario lo desea. El paquete **mice** detecta automáticamente la colinearidad en los datos para removerlos en el proceso de imputación.

En las tablas del apéndice B se muestran los resultados de las imputaciones generadas por el paquete **mice** para la variable *grado* y para la variable *etapa*. La primera columna representa un indicador para el paciente y las siguientes columnas representan las diferentes imputaciones. Cabe recordar que la variable *grado* tiene 160 datos faltantes mientras que la variable *etapa*

tiene 643 datos faltantes.

Una forma diferente de visualizar las imputaciones generadas es graficar las variables imputadas versus otra variable. Las figuras 5.8, 5.9, 5.10 y 5.11 muestran las diferentes imputaciones de la variable *etapa* y la variable *grado* versus la edad del paciente (*edad*) ó el tamaño del tumor(*tamao*), donde los círculos en color azul representan los valores observados y los círculos en color rojo representan los valores generados para cada imputación.

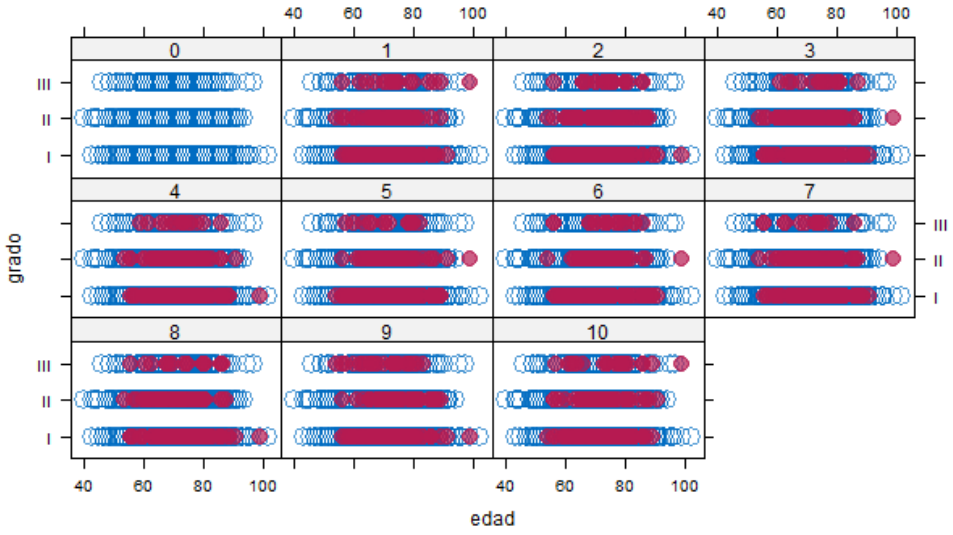

Figura 5.8: 10 imputaciones de la variable *grado* versus las variable *edad*, donde los puntos en azul representan los valores observados y los puntos en rojo los valores imputados.

El paquete estadístico R brinda otras alternativas diferentes para visualizar los datos imputados como se muestra en la figura 5.12 , donde se pueden observar nuevamente los valores imputados en color rojo y los observados en color azul de la variable *etapa* versus las variables *edad* y (*tratamiento*).

5.1.1. Análisis

Con las m imputaciones obtenidas por la técnica de múltiple imputación el siguiente paso es analizar las imputaciones con el modelo de interés, que en el presente trabajo es el modelo de Larson & Dinse para determinar la probabilidad de fallecer por cáncer de próstata dentro de un periodo utilizando la información recabada en el estudio. Pero antes de comenzar con el análisis de las imputaciones se procederá a definir algunos términos importantes y especificaciones del modelo para su uso posterior.

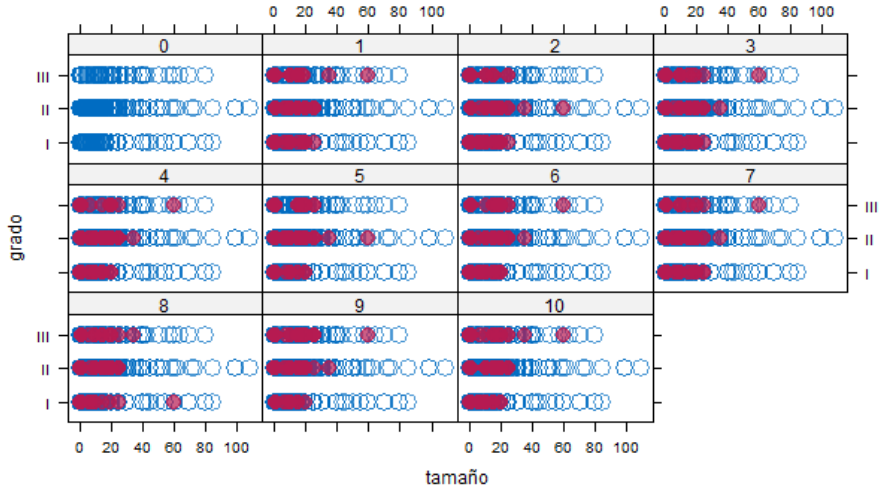

Figura 5.9: 10 imputaciones de la variable *grado* versus la variable *tamao*, donde los puntos en azul representan los valores observados y los puntos en rojo los valores imputados.

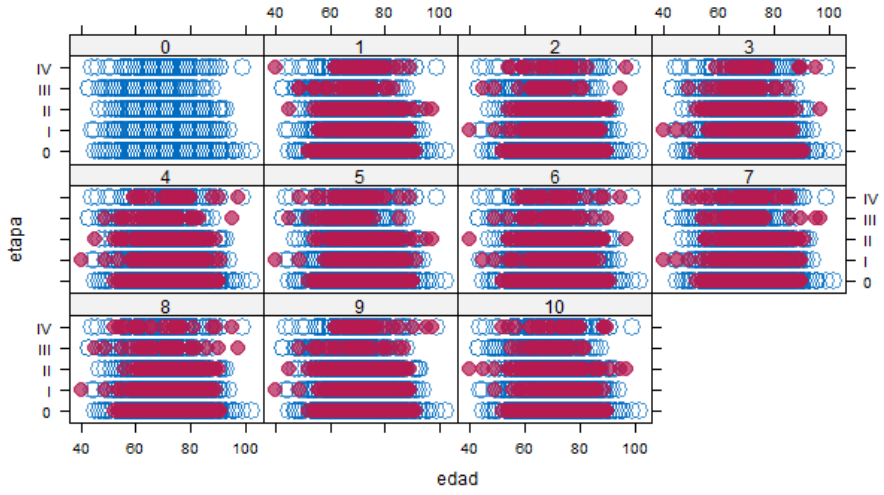

Figura 5.10: 10 imputaciones de la variable *etapa* versus la variable *edad*, donde los puntos en azul representan los valores observados y los puntos en rojo los valores imputados.

BIC(Bayesian Information Criterion)

El Bayesian Information Criterio (BIC) es un criterio para la elección de un modelo dentro de un conjunto finito de diferentes modelos, el cual es basado en la función de verosimilitud. Cuando se hace el ajuste de modelos, se puede incrementar la probabilidad mediante la adición de parámetros en el modelo, pero tal vez puede resultar un modelo sobre ajustado. Para resolver este problema, el BIC introduce un término de penalización para el número

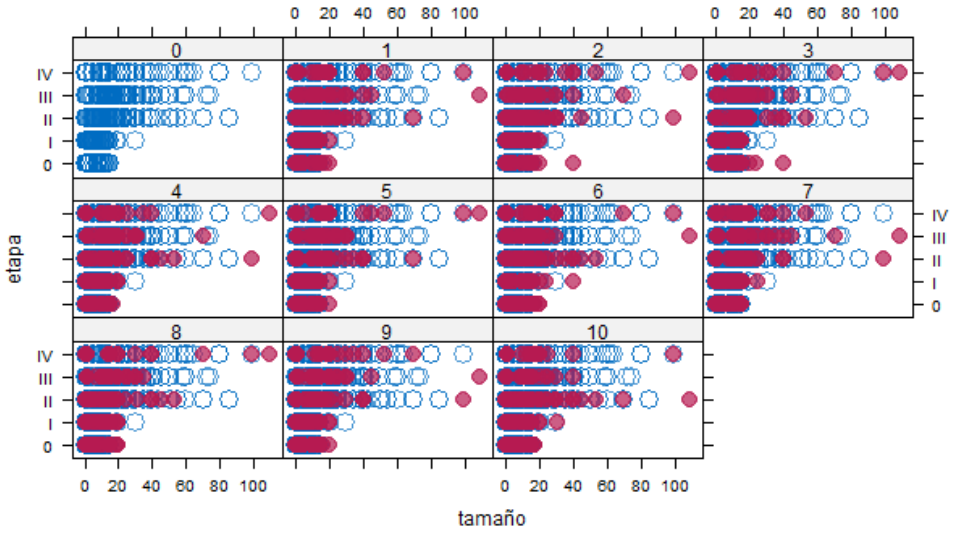

Figura 5.11: 10 imputaciones de la variable *etapa* versus la variable *tamao*, donde los puntos en azul representan los valores observados y los puntos en rojo los valores imputados.

de parámetros en el modelo. Matemáticamente el valor del BIC se puede calcular de la siguiente manera:

$$BIC = -2L(\widehat{\Theta}_p) + n_p log(n) \tag{5.4}$$

Donde $\widehat{\Theta}_p$ es la estimación de los parámetros de la máxima Verosimilitud del número de parámetros n_p, n es el número de datos y L es la función de verosimilitud. La idea del criterio es elegir el modelo con el BIC más pequeño.

Modelo de Larson & Dinse

Anteriormente ya se ha dado una descripción del modelo de Larson & Dinse, donde la probabilidad de supervivencia a la causa j está definida por la ecuación (3.16). Para los datos de Cáncer de Próstata sólo se tienen 2 causas de muerte, así $j = 1$ denotara la muerte por cáncer de próstata y $j = 2$ la muerte por otras causas.

De la ecuación (3.16)) se obtiene:

$$S_j(t) = S_{0j}(t)^{e^{\beta_j' \mathbf{x}}}, j = 1, 2. \tag{5.5}$$

Con $S_{0j}(t) = exp\left[-\Lambda_{0j}(t)\right]$ conocida como la función de supervivencia base condicional para la causa j, derivada del *modelo de riesgos proporcionales de Cox*.

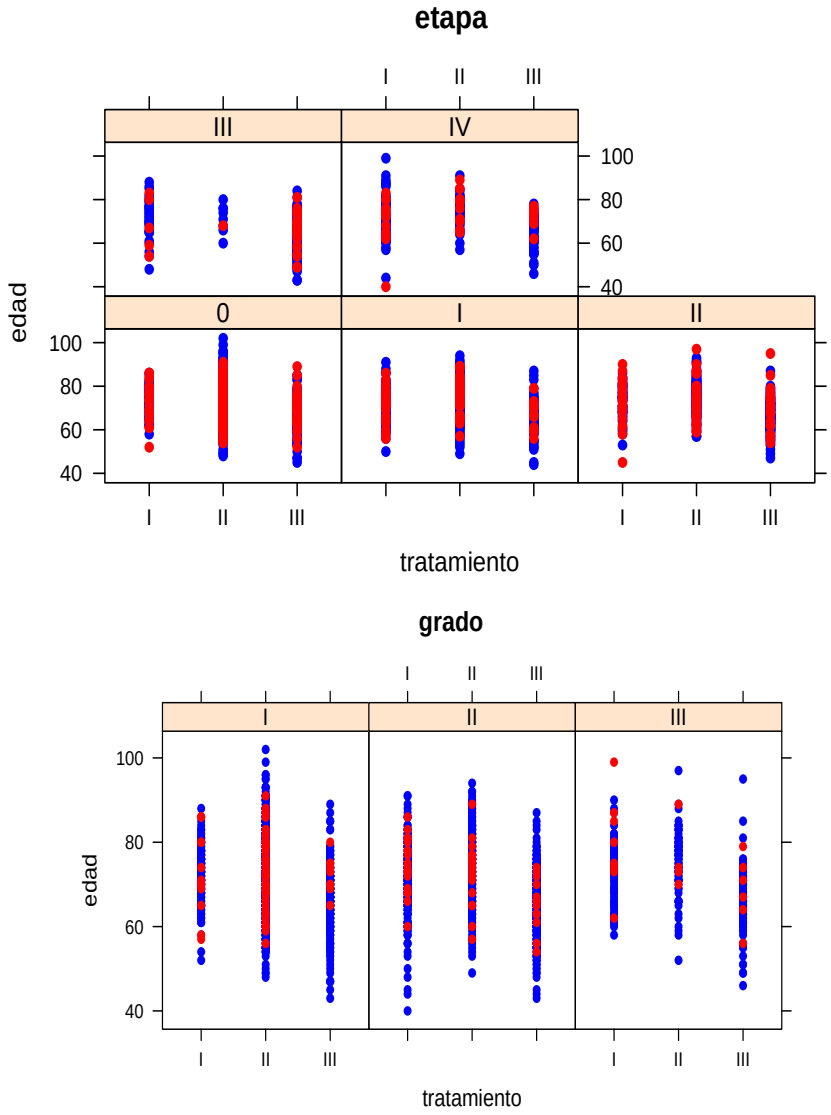

Figura 5.12: Visualización de las imputaciones para las variables *etapa* y *grado* por *tratamiento* y *edad*, donde los puntos en azul representan los valores observados y los puntos en rojo los valores imputados.

En el presente trabajo, la función de supervivencia base condicional para la causa j estará dada por el modelo Weibull como se puede ver en la ecuación (5.6), con parámetros desconocidos λ_j y γ_j ya que no se sabe cuál sea la tasa de muerte $h_{0j}(t)$ para las dos causas de muerte.

La importancia de utilizar el modelo Weibull radica en como sus parámetros interfieren con las tasas de muerte, en particular el valor de γ. Si $\gamma_j = 1$ la tasa de riesgo es constante en todo el tiempo. Si $\gamma_j > 1$ la tasa de riesgo se incrementa con el tiempo, mientras que para $\gamma_j < 1$ la tasa de riesgo decrece

con el tiempo. De esta manera los valores de λ_j y γ_j para las dos causas de muerte también se estimarán con la función de Máxima Verosimilitud del modelo de Larson & Dinse en la ecuación (5.1.1).

$$S_{0j}(t) = e^{-(\frac{t}{\lambda_j})^{\gamma_j}}, \quad \lambda_j, \gamma_j > 0 \tag{5.6}$$

con λ_j y γ_j los parámetros de forma y escala respectivamente para el riesgo $j = 1, 2$.

Por último, siguiendo la formulación del modelo de Larson & Dinse, las probabilidades p_j de que el paciente i muera por la causa j estarán dadas por un modelo logístico de la siguiente manera:

$$p_1 = \frac{e^{\alpha+\alpha_1'\mathbf{x}}}{1+e^{\alpha+\alpha_1'\mathbf{x}}}$$
$$p_2 = 1 - p_1$$

donde α y α_1', son los parámetros correspondientes a la intersección y a los coeficientes del vector con variables explicativas $\mathbf{x}$ para el paciente i.

En resumen la función de verosimilitud para el modelo de Larson & Dinse es el siguiente:

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

con $\Theta = (\alpha, \alpha_1, \beta_1, \beta_2, \lambda_1, \gamma_1, \lambda_2, \gamma_2)$ los parámetros a estimar.

Para la estimación de los coeficientes de las variables categóricas, el código del modelo de Larson & Dinse elaborado en el paquete estadístico R crea variables ficticias correspondientes a cada valor que estas pueden tomar, donde el valor base que se tomó es el primero en cada variable categórica y el coeficiente del valor base en cada variable categórica es igual a cero.

5.1.2. Análisis y resultados con las m bases datos completas

Una vez discutida la información sobre el BIC y las especificación del modelo de Larson & Dinse, el siguiente paso es analizar las m bases de datos generadas por el proceso de imputación múltiple una a una. La idea es utilizar este criterio en cada una de las bases de datos para la elección del mejor modelo en base a la función de Verosimilitud de Larson & Dinse .

El modelo seleccionado para cada una de las $m = 10$ bases de datos se ilustra en el Cuadro 5.4.

Cuadro 5.4: Modelo seleccionado cada imputación utilizando el criterio del BIC

IMPUTACIÓN 1	edad	etapa	tratamiento	civil	grado
IMPUTACIÓN 2	edad	etapa	tratamiento	civil	
IMPUTACIÓN 3	edad	etapa	tratamiento	civil	
IMPUTACIÓN 4	edad	etapa	tratamiento	civil	grado
IMPUTACIÓN 5	edad	etapa	tratamiento	civil	
IMPUTACIÓN 6	edad	etapa	tratamiento	civil	
IMPUTACIÓN 7	edad	etapa	tratamiento	civil	grado
IMPUTACIÓN 8	edad	etapa	tratamiento	civil	
IMPUTACIÓN 9	edad	etapa	tratamiento	civil	
IMPUTACIÓN 10	edad	etapa	tratamiento	civil	

Nótese que todos los modelos en las imputaciones son muy similares y la mayoría incluyen a las variables *edad,etapa,tratamiento,* y *civil* excepto por las imputaciones 1, 4 y 7 que además incluyen a la variable *grado*. Para la elección de las variables que se utilizarán en el modelo final Y. Vergouwe [18] y Van Buuren [23] proponen usar el método de *majoriy* (mayoría), el cual usa las variables que se encuentran en por lo menos la mitad de los modelos en cada una de las imputaciones

Así el modelo final quedará dado por las variables *edad*, *etapa*,*tratamiento* y *civil*. De modo que la siguiente parte es estimar los parámetros de las variables anteriores por el método de máxima verosimilitud del modelo de Larson & Dinse para cada imputación y aplicar la regla de Rubín para calcular las estimaciones de las combinaciones de estos parámetros.

En el Apéndice C se pueden encontrar las estimaciones de los parámetros del modelo y los de la función de supervivencia condicional base de las 10 imputaciones, mientras que en los Cuadros (5.6) y (5.5) se muestran los resultados obtenidos de la combinación de estos utilizando la regla de Rubín.

Cuadro 5.5: Parámetros de la función de supervivencia condicional base utilizando la regla de Rubín.

$\log(\lambda_1)$	5.25400	$\log(\gamma_1)$	0.38285
	(0.35450)		(0.01718)
$\log(\lambda_2)$	6.46555	$\log(\gamma_2)$	0.31200
	(0.04246)		(0.00172)

Una vez que se tienen la estimación de los coeficientes se puede calcular la probabilidad de que un paciente muera por Cáncer de próstata ó por causas ajenas a la enfermedad durante los proximos 20 años dadas las características con las que cuenta, utilizando la función de incidencia acumulativa definida como $p_j F_j(t)$, con $F_j(t) = 1 - S_j(t)$ y $S_j(t)$ en la ecuación (5.5). Por ejemplo supóngase que un paciente de 72 años fue diagnosticado con cáncer de próstata, que se encuentra en una etapa tempranamente avanzada, que además es casado y se sometió al tratamiento del tipo II, La probabilidad de que el paciente muera por Cáncer de Próstata se ilustra en las graficas de la Figura 5.13.

Cuadro 5.6: Valores de los coeficientes del modelo utilizando la regla de Rubín.

Variables	α	β_1	β_2
Intercepción	-1.23480	–	–
	(0.44081)	–	–
edad	-0.01250	0.05671	0.07541
	(0.00137)	(0.00451)	(0.00018)
etapa I	0.91439	0.19691	0.00853
	(0.06906)	(0.08511)	(0.01886)
etapa II	0.90490	-0.01484	-0.05111
	(0.09428)	(0.10407)	(0.02757)
etapa III	1.40438	0.71781	-0.00623
	(0.12301)	(0.12380)	(0.05846)
etapa IV	2.77832	1.15625	0.52365
	(0.11040)	(0.14540)	(0.13649)
civil otro	0.19147	0.44585	0.27647
	(0.03512)	(0.02996)	(0.00597)
tratamiento II	-0.22594	-0.04489	-0.02693
	(0.03218)	(0.06419)	(0.01039)
tratamiento III	-0.40809	-0.77548	-0.36622
	(0.08542)	(0.10454)	(0.02327)

De acuerdo a la Figura 5.13 se puede notar que la probabilidad de que el paciente muera de cáncer de próstata aumenta con respecto al tiempo, pero es muy pequeña con respecto a la probabilidad de que muera por otras causas.

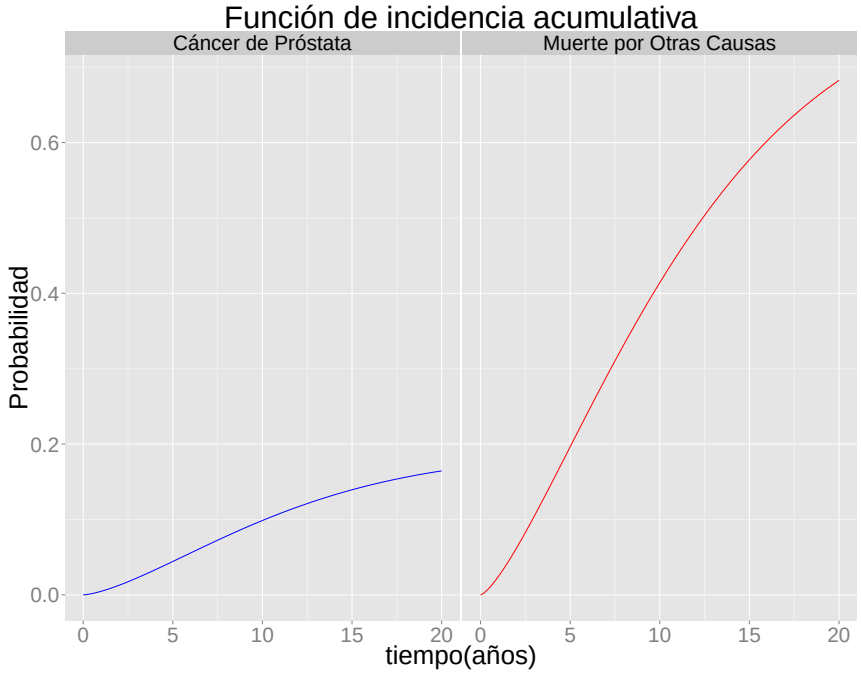

Figura 5.13: Probabilidad de que un paciente muera por cáncer de próstata ó por causas ajenas a la enfermedad dadas las variables: edad del paciente, tratamiento, la etapa del cáncer y el estatus marital.

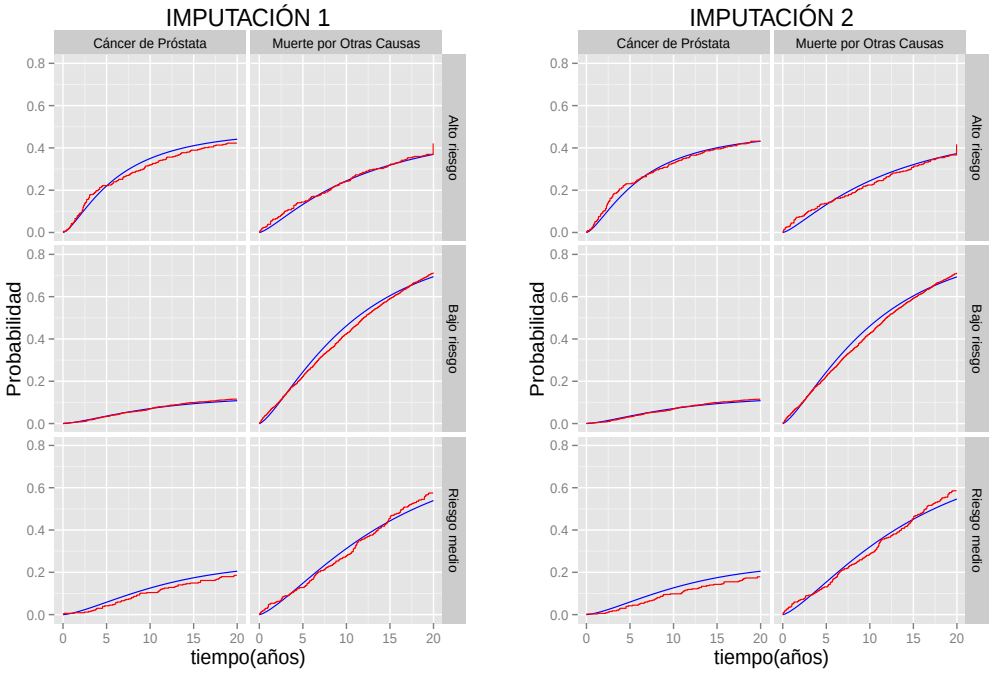

Figura 5.14: Función de incidencia acumulativa no paramétrica (rojo) y ajustada (azul) para la muerte por cáncer de próstata y otras causas en los tres grupos de riesgo.

El modelo propuesto es útil para investigar que se puede predecir la mortalidad por Cáncer de próstata y la de la muerte por otras causas, además de ser capaz de asignar diferentes grupos de riesgo a los pacientes que eventualmente podrían experimentar el evento de interés, esta asignación se realiza mediante la puntuación de Riesgo calculada por $Risk = \alpha + \alpha_1^{'}\mathbf{x}$, donde α y α_1 son los parámetros obtenidos de la combinación de Rubín en el Cuadro 5.6 y $\mathbf{x}$ las variables en la base de datos. Dado que el 16 % de los pacientes murieron de cáncer de próstata se tomo el 75 % de los valores más bajos de *Risk* como el grupo de bajo riesgo, a los siguientes 12,5 % valores más bajos de *Risk* se les asignó el grupo de riesgo medio, y los restantes valores del grupo de *Risk* fueron asignados al grupo de alto riesgo. Dados los grupos de riesgo, dos medidas importantes para la evaluación del ajuste del modelo son la *validación* y la *calibración*. Para la visualización de la *validación* se calculará la función de incidencia acumulativa en cada grupo de riesgo para los dos tipos de fracaso por cada una de las 10 imputaciones utilizando las estimaciones de los parámetros obtenidos por la regla de Rubín y su estimación no paramétrica. Las Figuras 5.14, 5.15 y 5.16 muestran los resultados de la validación en cada imputación.
Obsérvese que las funciones de incidencia acumulativas para la muerte por cáncer de próstata y la muerte por otras causas del modelo ajustado son muy similares a las no paramétricas para los diferentes grupos de riesgo en cada imputación, lo que sugiere que la clasificación es correcta y el modelo es adecuado. Calculando el promedio de las 10 imputaciones para cada grupo de riesgo de los valores de la función de incidencia acumulativa ajustada y de la estimación no paramétrica, se obtiene la figura(5.17), y de la misma manera se puede notar que las gráficas son muy similares.
Para la calibración Kattan MW [41] propuso un grafico para evaluar el modelo ajustado. El gráfico propuesto utiliza la estimación no paramétrica de la función de incidencia acumulativa de las dos causas de muerte para los diferentes grupos de riesgo contra con la media de la función de incidencia acumulativa (CIF) para la causa j de muerte $\frac{1}{d}\sum_{j=1}^{d} p_j F_j(t, x_i)$, donde d es el número de sujetos en el grupo de riesgo y x_i es el vector de variables explicativas para el i-ésimo paciente. Si el modelo es bueno, el gráfico resultante se debe de parecer a una línea recta con pendiente 1.

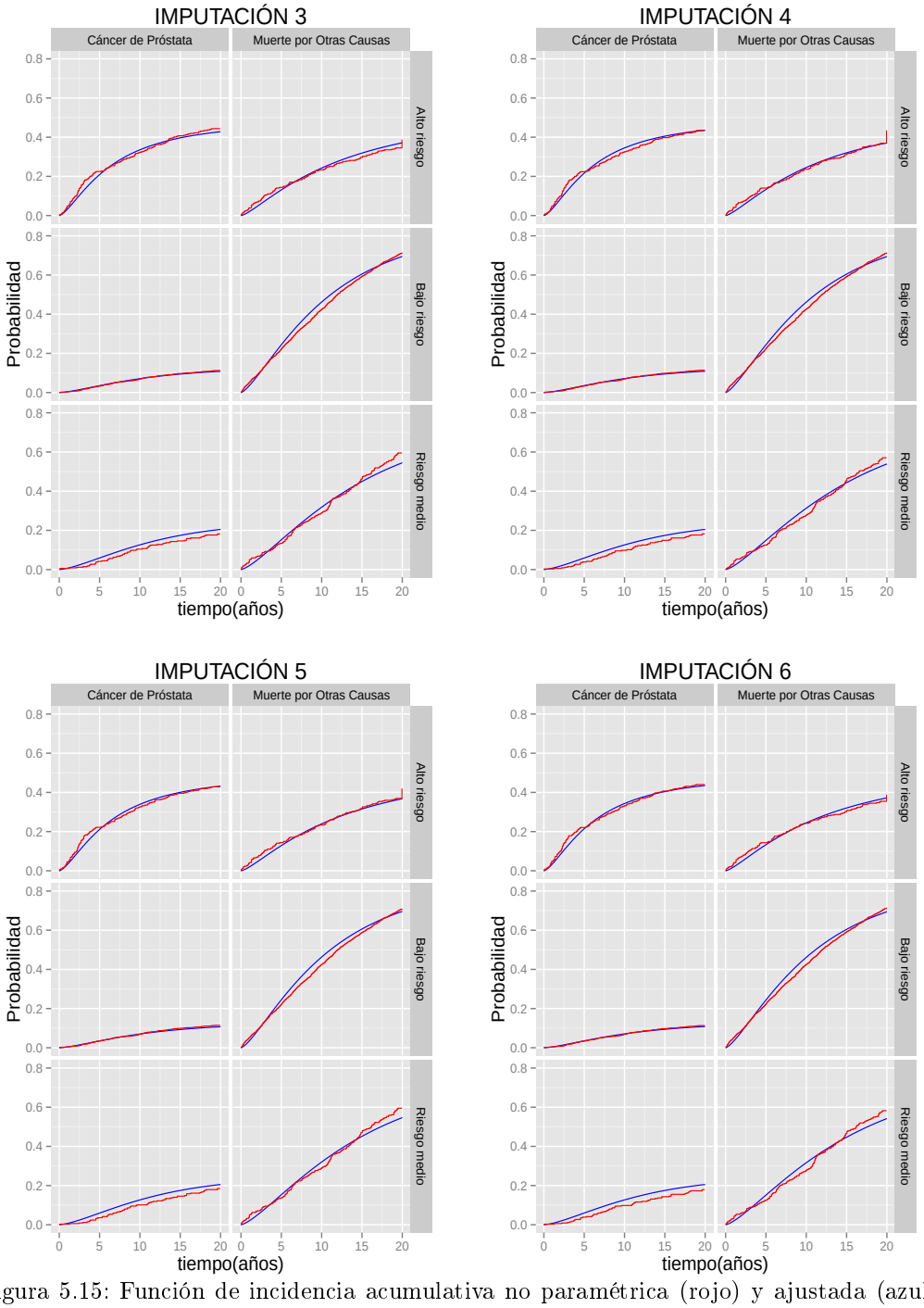

Figura 5.15: Función de incidencia acumulativa no paramétrica (rojo) y ajustada (azul) para la muerte por cáncer de próstata y otras causas en los tres grupos de riesgo*(CONTINUACIÓN)*.

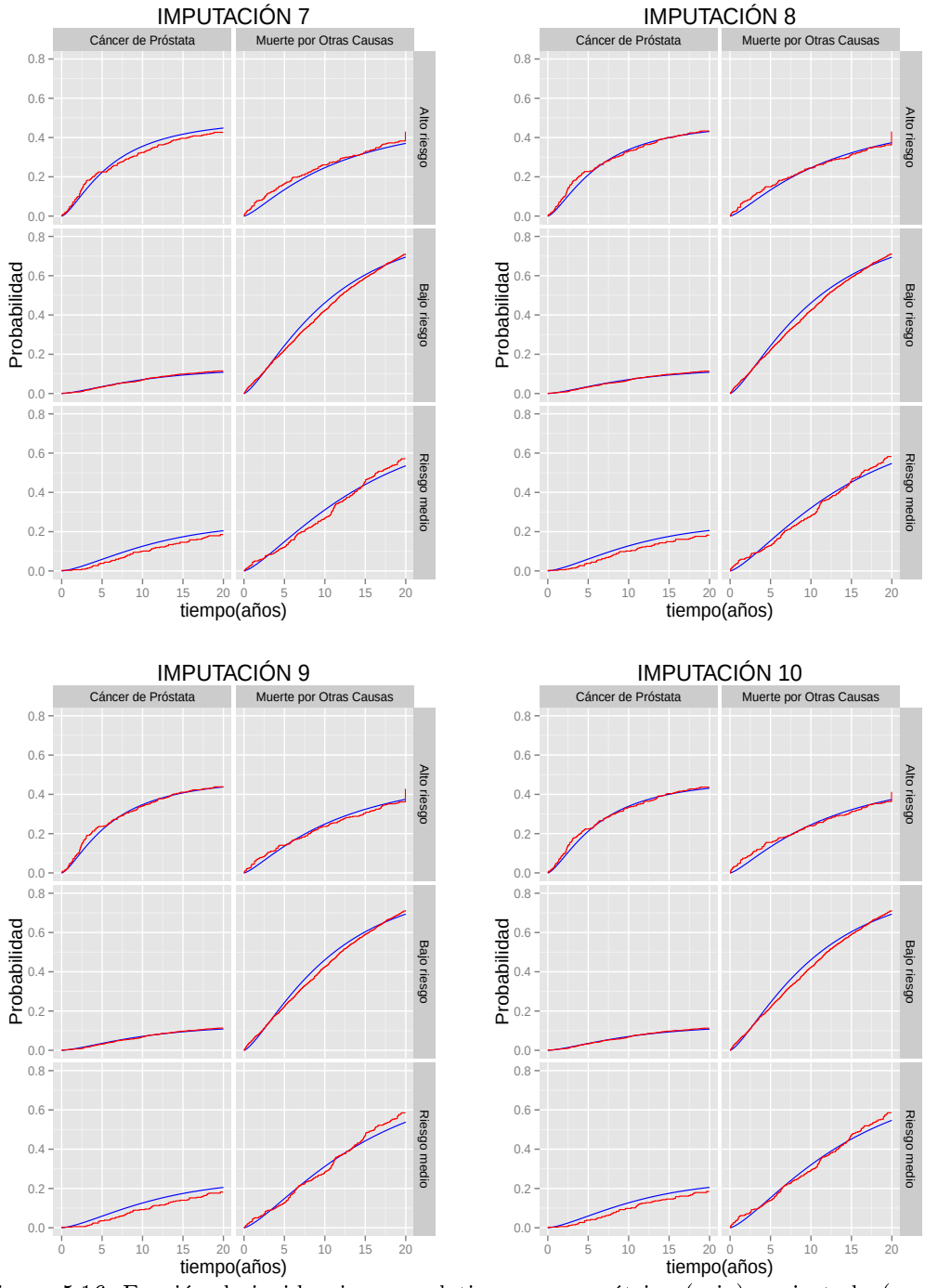

Figura 5.16: Función de incidencia acumulativa no paramétrica (rojo) y ajustada (azul) para la muerte por cáncer de próstata y otras causas en los tres grupos de riesgo*(CONTINUACIÓN)*.

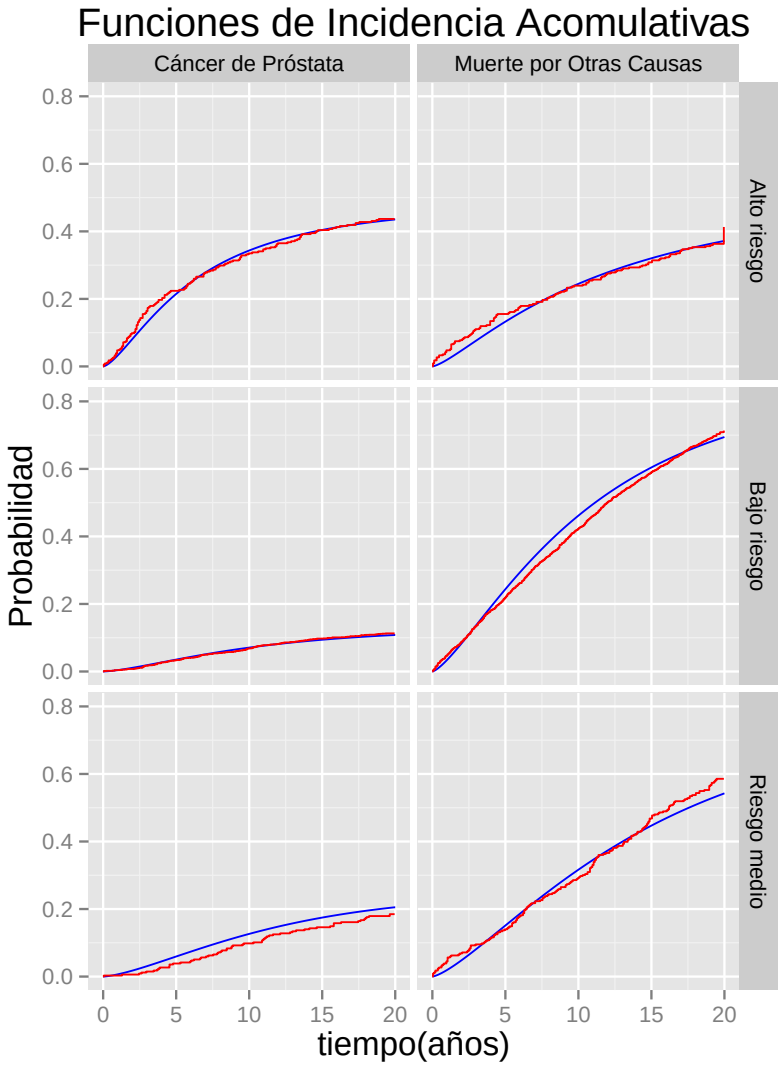

Figura 5.17: Promedio de las funciones de incidencia acumulativas no paramétrica (rojo) y ajustada(azul) de las imputaciones para los diferentes tipos de fracaso.

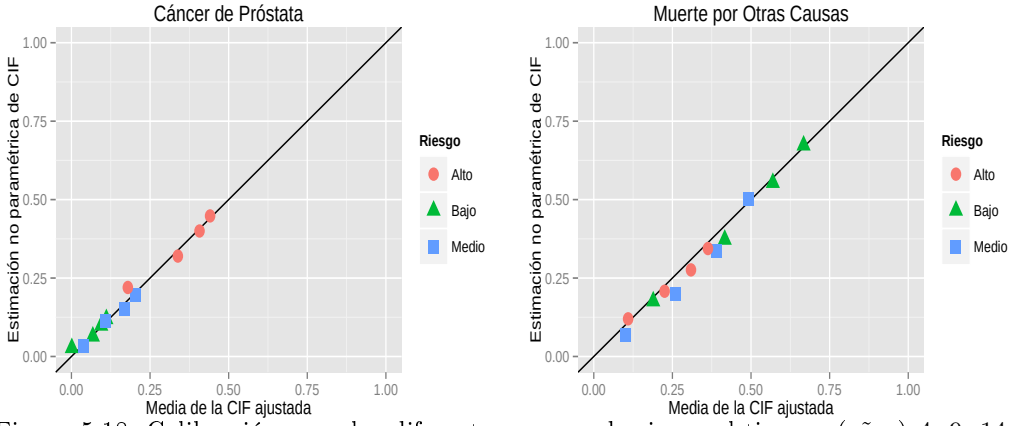

Figura 5.18: Calibración para los diferentes grupos de riesgo al tiempo (años) 4, 9, 14 y 19.

La Figura 5.18 ilustra la calibración de las funciones de incidencia acumulativa para la muerte de cáncer de próstata y muerte por otras causas correspondientes a los tiempos (años) 4, 9, 14 y 19 en los diferentes grupos riesgo. Las graficas de la Figura 5.18 muestran una tendencia a la línea recta por lo que el modelo es adecuado, de esta forma la función de incidencia acumulativa para los dos tipos de fracaso es muy cercana a las estimaciones de la función de incidencia acumulativa para todo los datos.

Conclusiones

En este trabajo se mostró un procedimiento estadístico para trabajar con estudios de análisis de supervivencia, en particular el poder determinar la probabilidad de que un paciente fallezca dentro de un periodo arbitrario de tiempo dado que fue diagnosticado con cáncer de próstata con variables explicativas que influyen en el estado de su salud.

El modelo de Larson & Dinse se utilizó para el análisis de estos datos por la existencia de diferentes tipos de muerte que pueden llegar a censurar el evento de interés. Además al no saber cuál era la tasa de muerte para las dos diferentes causas, la distribución Weibull en la función de supervivencia base condicional de este modelo jugó un papel importante por la flexibilidad que muestra su tasa de muerte para diferentes valores de sus parámetros de forma y escala. Los valores de escala en cada una de las funciones condicionales base resultaron mayores que 1, por lo que la tasa de riesgo de morir por alguna de las dos causas aumenta con respecto del tiempo, en particular es lo que se esperaba en el caso del cáncer de próstata.

Las variables que influyen en la probabilidad de morir por alguna de las causas en el modelo propuesto fueron la edad del paciente, la etapa del tumor, estado civil (*marital*) y el tipo de tratamiento de acuerdo a la selección de variables utilizando el criterio del BIC.

Como se puede ver en las diferentes gráficas de las funciones de incidencia acumulativas la probabilidad de morir por cáncer de próstata es pequeña con respecto a la probabilidad de morir por otras causas, esto se debe generalmente a que la esperanza de vida en los hombres estadounidenses es de 74 a 77 años y el cáncer de próstata ocurre en los hombres de edad avanzada donde habitualmente dos tercios de los casos de cáncer de próstata se diagnostican a hombres de 60 años o más, aunque muy pocas veces se presenta a edades menores de 40 años. En la Figura 5.19 se ilustran las edades de los pacientes de los datos utilizados en este estudio al momento del diagnóstico.

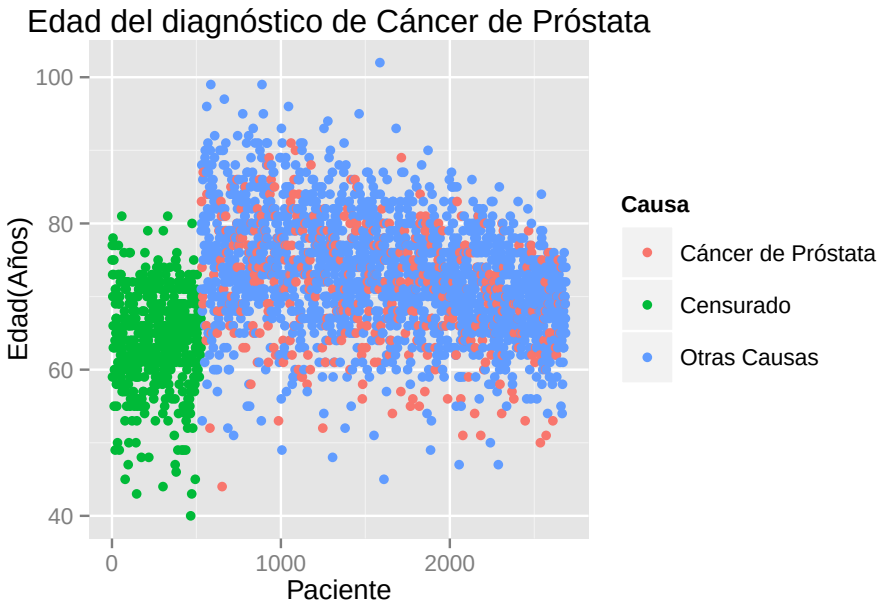

Figura 5.19: Edad del paciente al momento del diagnostico por Cáncer de Próstata y la causa de su muerte posterior.

En la actualidad el cáncer de próstata es considerado como una enfermedad letal aunque la velocidad del desarrollo del tumor cancerígeno es lenta con respecto al tiempo, de aquí la importancia de que la variable grado y el tamaño del tumor no sean importantes en el modelo, por que el 59 % de los datos corresponden al grado I, es decir que el tumor es diferenciable por lo que tiende a crecer y propagarse muy lentamente en contraste con el no diferenciable correspondiente al grado III que afecta al 10 % de los pacientes en este estudio.

La técnica de múltiple imputación jugó un papel trascendente en la presente investigación ya que gracias a esta se logro manejar las dos variables *etapa* y *grado* con datos faltantes para poderlas introducir en el análisis posterior del modelo propuesto. Tal vez la única desventaja de esta técnica radica en la aplicación computacional, y no por la creación de las imputaciones si no por el tiempo que utiliza el programa elaborado en **R** para el análisis posterior en el proceso de la selección de las variables en el modelo.

Al final el objetivo fue logrado con la propuesta del modelo de Larson & Dinse y las diferentes probabilidades de morir por las dos causas j en este estudio quedan de la siguiente manera:

$$P\{el\ paciente\ i\ muera\ por\ la\ causas\ j\} = p_j(1 - S_{0j}(t)^{e^{\beta_j' x_i}})$$

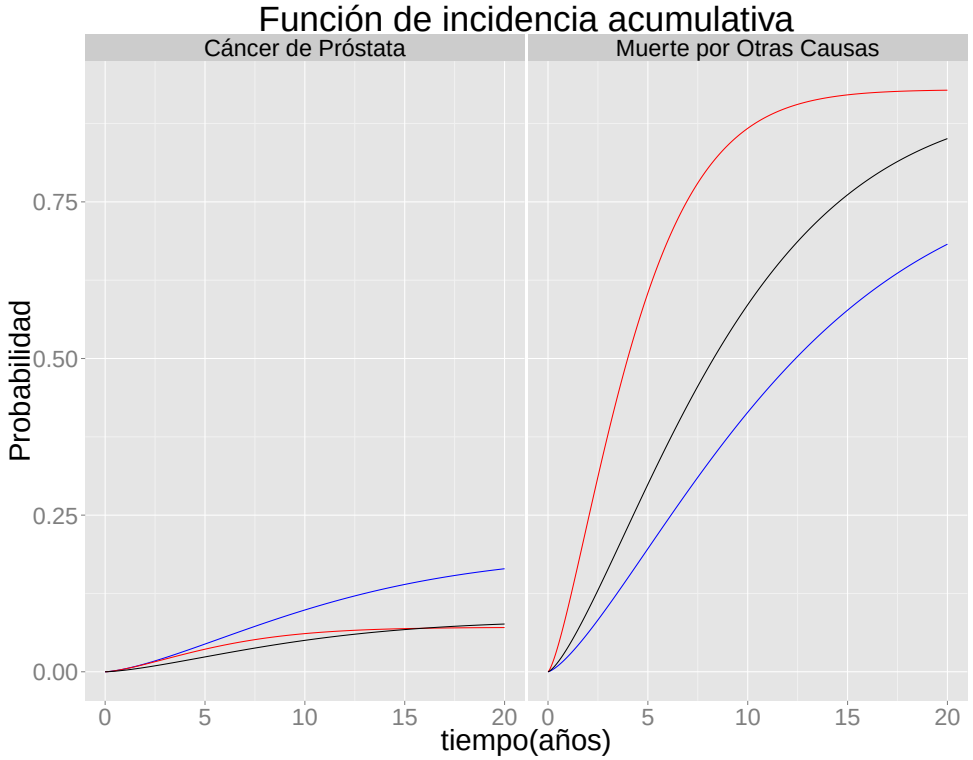

Figura 5.20: Probabilidad de morir por las dos diferentes causas para tres pacientes con características diferentes.

con $S_{0j}(t) = e^{-(\frac{t}{\lambda_j})^{\gamma_j}}$, $p_1 = \frac{e^{\alpha+\alpha_1' x_i}}{1+e^{\alpha+\alpha_1' x_i}}$, $p_2 = 1 - p_1$, los parametros (α, α_1, β_1, β_2, λ_1, γ_1, λ_2, γ_2) del cuadro (5.6) y x_i las variables explicativas del paciente i. De esta manera se muestran las funciones de incidencia acumulativas en la figura (5.20) de los dos tipos de causas de muerte para el modelo ajustado de Larson & Dinse a 3 diferentes pacientes.

El siguiente análisis es sobre un caso en particular, para eso supóngase que un paciente de 63 años es diagnósticado con cáncer de próstata, al momento del diagnostico el paciente se encuentra en la etapa I, el grado del tumor es moderadamente diferenciable y además es casado. Ahora la cuestión es ¿Cuál de los tres tratamientos es el adecuado para este paciente?, la respuesta la ilutra la Figura 5.21. Se puede observar que la linea en color negro correspondiente al tratamiento III (cirugía radical) es la más adecuada ya que muestra la menor probabilidad de morir por cáncer de próstata que las correspondientes al tratamiento I (linea en color rojo) y tratamiento II (linea en color azul).

La Figura 5.22 muestra las diferentes probabilidades del mismo paciente del ejemplo aplicándole el tratamiento III pero a diferentes valores de la

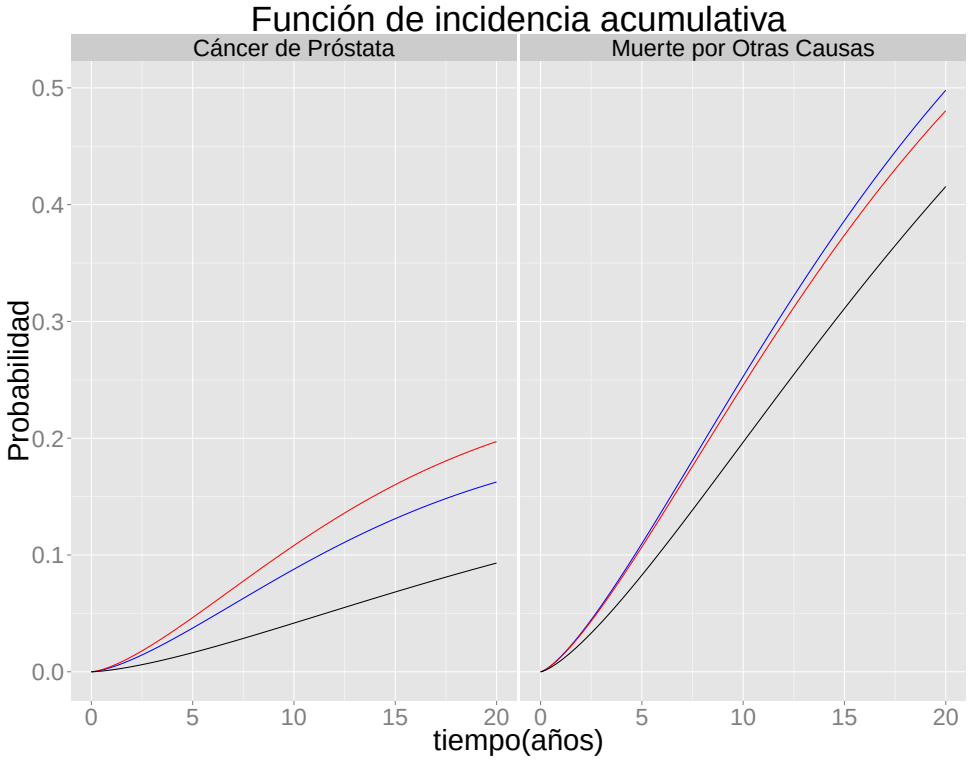

Figura 5.21: Las diferentes probabilidades de morir por una de las dos causas de un paciente si se somete al tratamiento I (líneas en rojo), al tratamiento II (líneas en azul) y al tratamiento III (líneas en color negro).

variable edad, si es que se hubiera detectado antes o después el cáncer de próstata a la fecha que se hizó el diagnóstico. La línea en color azul representa el diagnostico a la edad de 45 años, la línea en color rojo representa el diagnóstico ala edad de 63 y la línea en color negro representa el diagnóstico a la edad de 73 años. Nótese que la probabilidad de de morir por una de las dos causas aumenta con respecto al tiempo pero aumenta más rápido la probabilidad de morir por causas ajenas al cáncer y esto se debe a que el desarrollo del cáncer de próstata con respecto al tiempo es lento.

Por último la Figura 5.23 muestra las probabilidades de morir por Cáncer de próstata (líneas sólidas) o por otras causas (líneas punteadas) de las diferentes combinaciones de los valores en las variables que resultaron significantes en la base de datos del SEER. Obsérvese que en el caso del cáncer próstata el mejor escenario se tiene cuando el paciente está en la etapa IS, es casado y además se sometió al tratamiento III, mientras que el peor escenario se tiene cuando el paciente se encuentra en la etapa IV, tiene un estado civil distinto al casado y además se sometió al tratamiento I. La probabilidad más grande

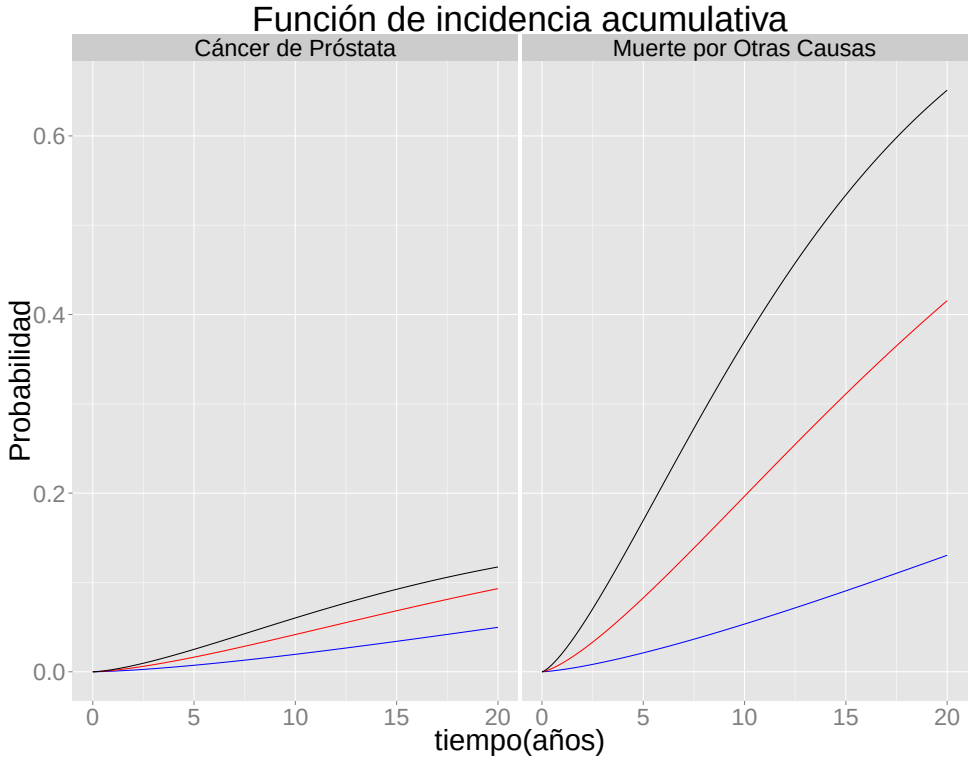

Figura 5.22: Gráfica correspondiente al mismo ejemplo del gráfico anterior, tomando el tratamiento que mejor se ajusto (III) aunque para diferentes edades del paciente.

de que un paciente muera por causas ajenas al cáncer de próstata se da cuando un paciente se encuentra en la etapa I, su estado civil es distinto al casado y se somete al tratamiento II (aunque es muy parecida la probabilidad si se somete al tratamiento I), aunque es menos probable que un paciente muera por causas ajenas al cáncer de próstata si esté se encuentra en la etapa IV, su estado civil es distinto al casado y además se somete al tratamiento I. En general para cada paciente se puede determinar la probabilidad de que muera por cualquiera de las dos causas una vez que se le haya diagnosticado el cáncer próstata.

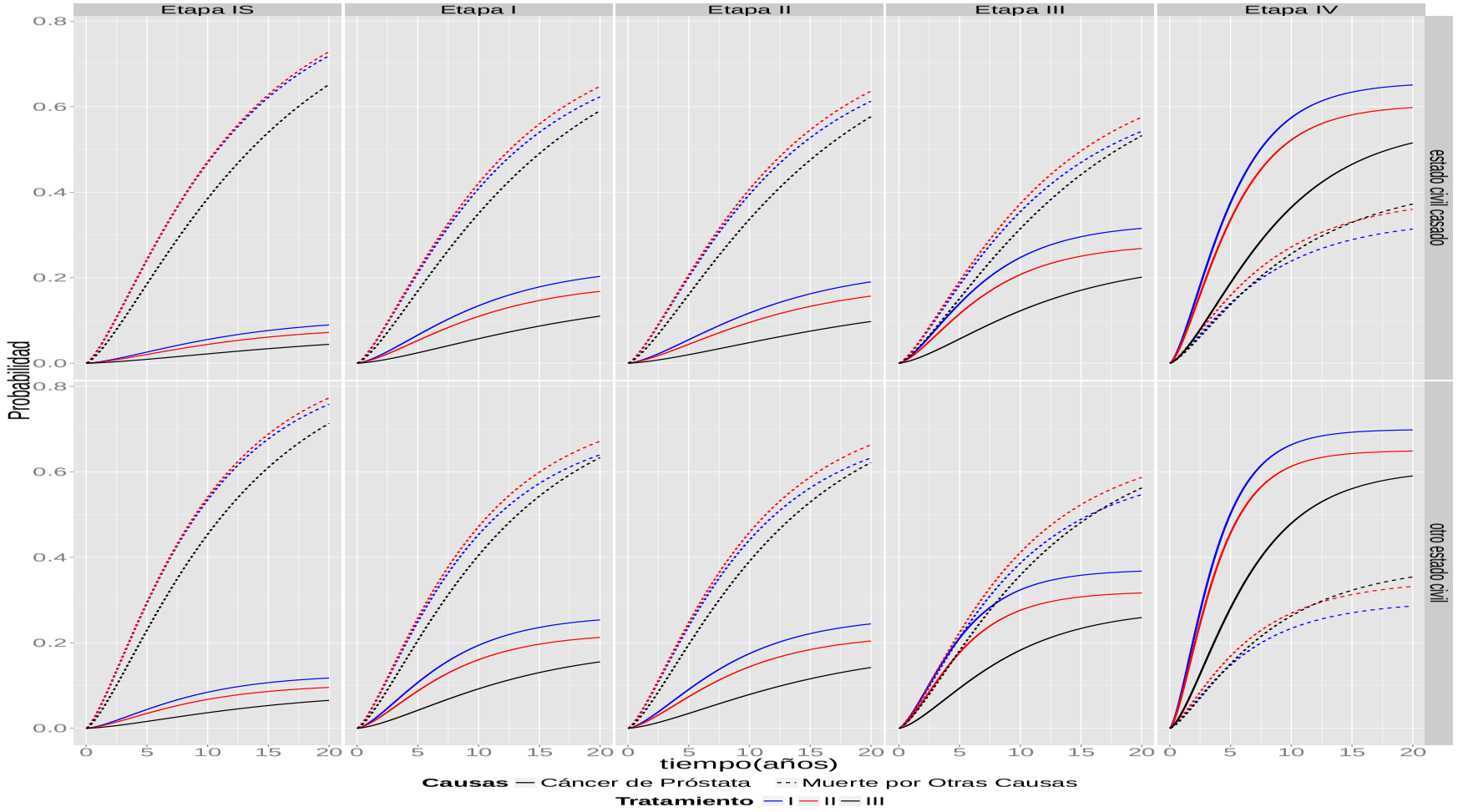

Figura 5.23: Probabilidad de morir por alguna de las distintas causas de acuerdo a las combinaciones de las variables que resultaron significativas en el análisis.

Apéndice

Apéndice A

Programas en el paquete estadístico R project

A.1. Función de máxima verosimilitud del modelo de Larson & Dinse

```
MCR <- 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]
    Vij <- matriz.S[,-1]
  }
  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]))
      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)))
    }
  }
  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
}
```

A.2. Estimación de los parámetros del modelo de Larson & Dinse

```
###Nombre del archivo
pc.dat <- read.table(file="pc.txt")

# La función objetivo del modelo de Larson & Dinse:
source(file="CRfunction.txt")

formula<-c("surgery+stage+grade+size.gp+age.gp+marital.gp")

formula<-as.formula(paste("~",formula))
print(formula)
###Evaluar formula y cada resultado guardarlo en un vector de
   nombre a
# Primero calcula el número de coeficientes en cada modelo:
n.par <- dim(model.matrix(formula, data=pc.cc))[2] -1
##Aplicar el modelo para maximizar la funcion de Larson y Dinse
system.time(
  mle.1 <- nlm(MCR,
               c(-1.360354, rep(0,n.par), rep(0,n.par), rep(0,n.
                  par),
                 2.453349, 2.646037, 0.01, -0.01),
               tipo=pc.cc$status, Ti=pc.cc$time,
               matriz.p= model.matrix(formula, data=pc.cc),
               matriz.S= model.matrix(formula, data=pc.cc),
               hessian=T, gradtol = 1e-3, iterlim = 500))

#Estimaciones
mle.1$estimate

##Estimadores
estima<-data.frame(estimate=mle.1$estimate)

delta<-estima[1:(n.par+1),]
beta1<-estima[(n.par+2):(2*n.par+1),]
beta2<-estima[(2*n.par+2):(3*n.par+1),]
logescala<-estima[(3*n.par+2):(3*n.par+3),]
logforma<-estima[(3*n.par+4):(3*n.par+5),]
```

Apéndice B

Imputaciones de las variables con datos faltantes en el estudio

Imputaciones de la variable ***grado***

No. Paciente	1	2	3	4	5	6	7	8	9	10
40	I	II	II	II	I	II	I	III	II	II
46	II	I	I	II	II	I	I	II	II	I
68	II	I	I	I	I	I	I	I	I	I
72	I	I	I	I	I	I	I	I	I	II
97	I	I	I	I	I	I	II	II	I	II
156	II	II	II	I	II	I	II	II	II	II
168	I	I	I	I	I	I	I	I	I	I
170	I	I	I	I	I	I	III	I	II	I
176	I	I	I	I	I	I	I	I	I	I
184	I	I	I	II	II	I	II	II	III	I
199	II	II	I	III	II	I	I	I	III	I
212	II	I	II	I	I	I	I	I	II	I
243	I	I	I	I	I	I	I	I	I	I
288	II	I	I	II	I	I	I	I	I	II
293	I	II	I	I	II	I	I	I	I	I
296	II	I	II	I	II	II	II	III	II	I
309	II	II	II	I	I	II	III	III	I	II
315	I	I	I	II	I	I	I	I	I	I
322	I	II	I	I	I	I	I	I	I	I
366	II	I	I	I	I	I	II	I	I	II
421	I	II	II	II	III	II	II	II	II	I
464	II	II	II	II	I	II	II	II	III	I
482	I	I	II	I	I	I	II	I	I	I
484	I	I	I	I	I	II	I	II	I	I
489	I	I	II	II	II	I	I	I	II	I
524	II	I	I	II	II	II	II	II	III	II
525	II	III	II	III	III	II	I	II	III	II
538	I	I	I	I	II	I	I	I	I	I
544	III	I	II	II	I	I	III	III	II	II
549	I	II	II	II	I	I	I	III	II	I
557	II	III	I	I	II	II	II	II	III	I
570	I	II	II	I	II	I	I	I	I	I
585	III	I	II	I	II	II	II	I	I	III
603	III	II	II	II	II	II	III	II	III	I
625	I	II	II	I	I	I	I	II	I	I
626	I	II	I	I	II	III	II	I	I	II
645	I	I	I	I	II	III	I	I	III	I
647	II	I	I	I	I	I	I	I	II	III
656	II	II	III	II	II	II	II	I	I	III
664	II	III	III	II	II	I	I	III	III	I
669	I	II	I	I	I	I	I	II	II	II
687	I	I	I	II	III	III	II	II	II	II
695	II	II	II	I	I	I	II	II	I	I
710	II	II	I	II	I	III	II	I	I	I
713	I	I	I	II	I	I	I	II	I	I
725	III	II	I	I	II	I	II	I	II	II
729	I	I	I	III	II	III	II	II	II	I
734	I	I	II	I	I	III	I	I	III	II
737	I	I	I	II	I	I	I	I	I	I
746	I	III	I	II	I	I	I	I	II	II
764	II	II	II	III	II	II	II	II	III	II
796	I	I	I	I	I	II	I	I	I	I
797	I	I	I	II	II	I	I	I	I	II
799	III	II	III	I	I	II	II	III	II	II
808	I	I	I	I	I	II	I	I	I	I
818	I	I	I	I	I	I	I	I	I	I
840	I	I	I	II	I	I	III	I	I	I
852	III	I	I	II	I	II	I	I	II	I
872	II	II	III	II	II	II	II	II	II	II
874	I	I	II	III	I	II	II	III	II	III
878	I	II	I	I	I	I	I	I	III	I
882	II	I	II	II	I	I	II	II	II	II
883	I	II	III	III	III	II	I	III	III	II
887	II	I	I	I	I	I	I	I	I	I
892	I	III	I	I	I	I	I	III	II	III
931	I	II	I	I	I	I	I	I	II	III
932	II	I	I	I	I	I	I	I	I	I
940	I	I	I	I	I	I	I	I	I	III
955	I	II	I	I	II	I	I	I	II	II
963	I	III	III	III	II	II	I	II	II	II
964	I	III	I	I	I	I	I	I	I	I
966	II	I	III	III	II	I	I	I	I	II
972	III	II	III	III	II	III	III	I	III	III
995	I	I	II	I	I	I	I	II	I	I
1019	I	I	I	II	I	I	I	I	I	I
1020	I	I	I	II	II	I	II	I	I	I
1035	I	I	I	I	I	I	I	I	I	I
1059	I	III	II	II	II	II	I	I	III	I
1113	I	II	I	I	I	I	I	II	II	I
1133	II	III	II	II	II	II	III	II	II	I

No. Paciente	1	2	3	4	5	6	7	8	9	10
1176	III	I	I	I	I	II	II	II	I	I
1184	I	II	I	I	I	II	II	I	II	I
1203	II	II	I	I	I	III	I	I	I	II
1206	I	II	II	I	III	II	I	I	II	II
1224	I	I	I	II	I	II	II	I	I	II
1230	II	III	II	I	III	III	II	III	I	II
1267	I	I	I	I	I	I	I	I	I	I
1282	II	I	I	I	II	I	I	I	I	I
1288	III	II	III	II	II	I	II	II	III	III
1316	I	III	II	II	I	III	II	II	II	II
1318	III	II	II	II	II	III	I	I	II	II
1330	II	II	III	II	I	II	III	III	II	II
1337	I	I	II	I	I	II	II	I	I	I
1338	I	I	I	I	I	I	II	I	I	III
1347	II	II	II	II	I	II	I	II	III	III
1355	I	I	III	I	II	II	I	II	II	I
1376	II	II	III	I	I	II	I	II	II	II
1389	I	III	I	I	I	I	I	I	I	II
1397	I	I	I	II	II	III	II	III	II	II
1406	I	I	I	I	II	II	I	I	I	II
1458	II	II	III	II	II	II	I	II	III	III
1464	I	I	I	II	I	I	I	I	II	I
1481	II	II	II	II	III	II	II	II	III	III
1482	II	III	II	II	II	III	III	I	III	II
1511	II	I	II	I	III	I	I	II	III	II
1529	I	I	II	II	II	I	I	I	I	I
1533	II	I	I	I	I	I	I	I	I	I
1612	II	II	II	III	III	III	III	II	II	III
1674	II	I	I	II	I	I	I	I	I	I
1675	II	I	I	I	I	I	I	II	I	I
1702	I	II	II	II	III	I	III	III	I	II
1713	III	I	I	I	I	I	I	I	II	III
1758	II	II	II	II	I	II	I	I	I	II
1770	I	I	I	II	I	I	I	I	II	I
1773	II	I	I	II	II	II	II	I	I	I
1787	I	I	III	I	II	II	I	I	I	I
1796	I	I	II	I	III	II	I	I	III	I
1800	III	II	II	I	III	I	II	I	I	I
1825	I	III	II	I	III	II	II	I	I	I
1852	I	I	II	I	III	I	I	II	III	III
1858	III	II	II	II	II	II	II	III	II	III
1874	I	III	II	II	I	I	I	I	I	II
1875	I	I	I	II	I	I	II	I	I	I
1877	I	I	I	III	I	I	I	I	I	I
1883	II	I	I	I	II	II	II	I	I	II
1905	I	I	I	I	I	I	I	I	II	III
1938	I	I	III	I	I	II	III	II	II	II
1963	I	I	I	I	I	I	I	I	I	I
1965	II	II	II	I	I	III	I	I	I	I
1974	I	I	I	I	I	II	I	I	I	I
1980	II	I	II	III	I	II	II	I	I	I
1997	I	III	II	I	I	I	I	II	I	I
2025	I	II	I	II	I	I	I	I	I	I
2037	III	II	II	III	II	II	II	II	III	III
2065	III	II	I	II	I	III	III	III	II	II
2122	II	II	II	II	II	III	II	II	II	I
2128	I	II	I	I	II	II	II	I	II	I
2161	II	I	I	II	III	I	I	II	II	II
2171	I	I	I	I	II	II	I	II	I	I
2176	I	I	II	II	I	I	II	I	II	I
2186	II	II	II	III	II	II	II	II	II	I
2204	I	I	I	I	I	III	I	II	II	I
2211	II	I	III	I	II	I	I	II	I	III
2314	I	II	III	II	II	I	I	II	I	III
2319	III	II	II	II	II	III	III	II	I	III
2334	I	I	I	I	I	I	II	I	I	I
2355	I	I	I	II	I	I	I	II	III	I
2402	III	III	II	II	II	II	II	I	I	II
2434	I	II	II	I	I	I	I	I	II	I
2452	I	I	II	II	I	I	I	II	I	I
2456	I	I	I	I	I	I	II	I	I	II
2512	II	II	I	I	I	II	I	I	I	I
2532	II	I	I	II	I	II	I	I	I	I
2549	I	I	I	I	I	I	II	I	I	I
2559	I	I	II	I	I	I	II	I	I	I
2595	III	II	II	II	I	II	I	II	I	II
2608	II	II	II	III	II	II	II	II	II	II
2642	II	III	II	II	I	II	II	III	II	II
2666	I	I	I	I	I	I	I	I	I	I
2681	I	I	I	I	I	I	I	I	II	I

Imputaciones de la variable ***etapa***

No. Paciente	1	2	3	4	5	6	7	8	9	10
2	IV	IV	IV	0	IV	II	IV	II	IV	II
5	III	I	I	IV	I	IV	IV	I	IV	IV
6	0	0	0	0	0	0	0	II	0	0
10	III	IV	IV	IV	IV	III	III	IV	III	II
11	II	0	0	II	0	0	0	II	0	0
13	0	0	0	0	0	0	II	0	0	0
20	II	0	0	0	0	0	II	0	II	0
22	0	II	0	0	0	II	0	II	0	0
28	0	0	0	0	0	0	0	0	0	0
40	II	I	I	I	0	I	II	I	II	I
44	IV	I	III	II	I	III	III	II	I	I
46	I	III	0	I	I	II	0	I	II	0
56	III	III	I	I	III	IV	III	III	IV	IV
68	I	0	0	0	0	0	0	0	II	0
71	II	0	0	0	0	0	0	0	0	0
72	0	0	II	0	0	0	II	II	II	I
86	0	II	II	0	0	0	0	II	0	0
91	II	0	0	0	0	0	0	0	0	0
93	I	I	II	III	II	I	III	III	III	II
94	II	II	III	III	II	I	I	III	I	I
97	II	0	0	0	0	II	IV	I	0	III
102	II	II	II	II	II	II	0	III	II	II
104	0	II	II	II	II	0	II	0	II	II
111	II	III	IV	III	III	IV	III	III	III	II
116	II	II	0	0	0	0	0	II	0	0
121	0	0	0	II	0	0	0	0	0	0
122	0	0	0	0	0	0	0	0	0	0
124	II	0	0	0	0	0	II	0	0	0
126	0	0	0	0	0	0	0	II	0	0
131	0	0	0	II	0	0	II	0	0	0
132	0	0	0	0	II	II	0	0	0	II
134	0	0	0	II	0	0	0	0	0	0
137	I	I	I	III	III	I	I	I	I	II
138	0	III	0	II	0	0	0	II	0	0
156	III	III	IV	III	II	II	III	III	I	II
164	0	0	0	0	0	0	0	0	II	0
166	0	0	0	0	0	0	0	0	0	0
168	0	II	II	0	0	II	II	0	II	II
170	0	0	0	0	0	0	I	0	I	0
175	II	0	0	II	II	II	0	0	0	0
176	0	0	II	0	0	0	0	0	III	0
179	0	0	II	0	II	0	0	0	II	II
184	IV	0	II	I	I	III	I	I	I	0
193	0	0	0	II	0	II	0	0	0	0
195	0	0	II	0	II	II	0	0	0	0
199	I	I	0	I	I	II	II	0	I	II
211	0	0	0	0	0	0	0	0	0	0
212	I	0	I	0	0	0	0	0	I	0
220	III	0	0	0	0	0	0	0	0	II
230	0	0	0	IV	0	0	0	0	0	0
236	IV	II	III	III	III	I	I	II	0	III
245	0	0	0	II	0	0	0	0	0	0
252	0	0	II	0	0	0	0	II	0	0
253	0	0	0	0	0	0	0	0	0	IV
265	II	II	III	II	0	II	II	III	0	II
288	I	0	0	I	0	IV	0	0	0	I
291	II	II	0	0	0	0	0	0	0	0
293	0	I	0	0	I	0	0	0	0	0
315	0	II	0	I	0	0	0	0	0	II
318	II	0	0	0	II	0	0	0	II	0
322	II	I	0	0	0	IV	IV	0	0	0
324	0	0	0	0	0	0	0	0	II	II
328	I	I	I	I	I	III	I	I	III	I
329	0	0	0	II	0	II	0	0	0	0
330	0	0	0	0	0	0	0	0	0	0
331	II	II	II	0	0	0	II	0	0	0
336	II	II	0	0	0	0	0	0	II	0
343	0	0	II	II	IV	II	0	III	0	0
347	I	III	IV	IV	I	III	IV	III	III	III
357	III	I	II	I	I	II	II	I	I	I
366	I	II	0	0	0	0	I	0	II	II
373	II	0	II	III	II	0	II	II	II	0
375	0	0	0	II	0	0	IV	0	0	II
382	0	0	0	0	0	0	0	II	0	0
386	0	0	II	0	0	0	II	0	0	0
389	0	0	II	0	0	0	0	0	0	0
391	I	I	I	I	I	I	I	I	I	I
394	0	0	0	0	0	0	0	0	0	0
395	0	0	0	0	0	0	0	II	0	0
399	0	0	II	0	0	II	0	0	II	0

No. Paciente	1	2	3	4	5	6	7	8	9	10
400	0	II	0	0	0	0	0	0	0	II
402	0	II	0	0	0	0	0	II	0	II
404	I	I	I	I	I	I	I	I	III	I
406	II	0	0	II	0	0	II	0	0	0
426	III	IV	III	III	III	I	III	III	III	III
433	0	0	0	0	II	0	0	II	II	0
435	0	0	0	II	0	0	0	0	0	0
436	0	0	II	II	0	0	0	II	II	0
437	III	I	I	I	I	I	I	I	I	I
442	0	0	0	0	0	0	0	0	0	II
443	II	0	0	0	0	0	0	0	IV	0
450	0	0	0	0	0	0	0	0	0	0
451	II	II	0	II	0	0	0	0	0	0
457	0	0	0	0	0	0	0	0	0	0
464	III	II	II	I	0	I	I	I	III	II
466	IV	I	I	I	I	II	I	I	I	II
479	0	II	0	0	0	0	0	II	0	0
480	II	II	0	0	0	0	0	0	II	II
482	0	II	IV	0	0	II	0	I	0	0
484	0	II	0	II	0	II	0	I	0	0
486	0	II	II	0	0	0	0	II	0	0
489	0	0	I	I	I	0	II	0	II	0
498	IV	II	II	0	0	IV	0	0	II	0
503	0	0	0	II	0	0	0	0	0	0
509	I	II	III	II	III	III	II	I	I	III
511	III	IV	III	III	II	IV	III	I	III	III
513	I	IV	III	IV	II	I	III	II	I	I
514	III	III	I	III	III	III	IV	III	III	II
516	III	III	III	II	I	II	II	III	I	III
528	I	0	I	I	I	II	I	I	I	III
529	I	I	II	II	I	I	I	I	I	I
531	0	0	II	0	0	0	0	0	0	IV
536	0	II	0	0	0	II	II	0	0	0
538	0	II	0	0	I	0	0	0	0	0
543	0	0	0	0	0	0	0	0	0	0
544	II	0	I	I	0	0	I	II	II	I
545	II	0	III	II	0	0	IV	III	0	0
549	0	IV	II	I	0	0	0	IV	II	0
551	0	0	0	0	0	0	II	II	0	IV
556	0	0	0	II	0	0	II	0	II	II
559	III	IV	III	IV	IV	IV	IV	IV	IV	IV
563	I	II	I	I	I	I	I	I	I	I
565	0	II	0	0	0	0	0	0	0	0
566	0	0	0	0	0	0	0	0	0	0
567	0	0	0	0	0	0	0	0	II	0
570	0	I	I	0	I	0	0	II	0	0
573	I	I	I	IV	I	I	I	I	I	I
577	0	0	0	0	0	0	0	0	0	IV
579	0	0	0	0	II	0	0	0	0	0
580	0	0	II	0	0	0	IV	IV	0	IV
582	II	II	0	0	0	0	0	0	0	0
583	0	0	0	0	0	0	0	0	0	II
587	IV	IV	II	II	IV	II	IV	II	IV	II
589	0	II	0	0	II	0	II	II	II	0
596	0	0	0	0	0	0	0	0	0	0
605	0	II	0	0	0	0	IV	II	II	II
607	I	II	IV	I	II	II	I	I	I	IV
616	0	0	0	II	0	0	0	0	0	0
619	0	0	0	0	II	0	0	0	0	0
620	0	0	II	0	0	0	0	0	0	0
621	0	0	0	III	II	0	II	0	0	0
625	0	I	I	0	0	0	0	I	II	0
626	0	IV	IV	0	I	I	I	0	0	I
637	I	I	II	I	I	I	I	I	I	I
638	II	0	0	0	0	0	0	0	0	0
639	0	0	0	0	0	0	III	II	0	0
642	0	I	0	II	II	0	II	0	0	0
643	0	II	0	0	0	0	0	0	II	0
645	0	IV	0	III	I	IV	0	0	IV	0
647	I	0	0	0	II	II	0	0	I	III
649	0	0	0	0	0	0	0	0	0	0
656	I	I	I	I	I	I	I	0	0	II
660	0	II	0	0	0	II	0	IV	0	0
665	II	IV	II	IV	II	II	III	III	IV	II
677	III	II	III	III	III	IV	IV	IV	IV	III
679	0	0	0	0	0	0	0	0	II	0
682	0	0	0	0	0	IV	0	0	0	0
686	0	0	0	II	III	0	0	0	II	0
692	0	0	0	II	0	0	0	II	0	0
700	0	0	0	0	0	0	0	0	0	II

Imputaciones de la variable ***etapa*** (*Continuación*)

No. Paciente	1	2	3	4	5	6	7	8	9	10
701	I	II	I	I	I	I	I	I	I	I
702	0	0	0	0	0	0	0	0	0	0
710	I	I	0	I	0	II	III	0	0	0
728	II	III	III	III	III	III	I	IV	I	I
732	II	0	0	II	II	0	0	0	0	0
734	0	0	I	0	0	II	0	0	I	I
735	II	0	II	0	0	0	0	0	0	0
737	0	0	II	I	0	0	0	0	II	0
738	0	0	0	0	0	0	0	II	II	0
746	0	IV	0	I	0	0	0	0	I	I
747	0	0	0	0	0	II	0	0	II	0
774	II	III	IV	III	II	IV	III	IV	IV	II
778	0	0	0	0	0	0	0	0	0	0
779	0	0	0	0	0	0	0	0	II	0
782	I	I	I	II	I	I	II	I	II	I
783	0	0	0	0	0	0	0	0	0	0
796	0	0	0	II	0	II	0	0	0	0
797	0	II	0	I	I	0	II	0	0	II
799	II	I	I	0	0	I	II	I	III	II
808	0	0	0	0	0	I	0	0	0	II
810	III	II	II	I	II	III	II	II	II	III
812	0	0	0	II	0	0	0	0	II	0
813	I	I	I	I	I	IV	II	I	IV	I
814	0	0	0	0	0	0	0	0	0	II
816	0	0	0	0	IV	0	0	0	II	0
818	II	0	0	II	0	0	0	0	0	II
819	0	II	0	0	0	II	0	0	0	0
821	I	I	I	IV	I	I	I	I	I	II
823	IV	II	IV	II	IV	IV	II	IV	II	IV
829	II	II	0	0	0	0	0	II	II	0
833	IV	III	III	III	II	III	IV	III	IV	II
837	0	0	0	0	0	0	0	II	0	0
840	II	II	0	I	0	0	IV	0	0	0
842	0	0	0	0	0	0	0	0	II	0
846	0	0	0	0	0	II	0	0	0	0
852	IV	0	II	I	II	I	0	0	I	0
853	0	0	0	0	0	0	0	0	0	0
859	IV	IV	II	II	IV	II	II	IV	IV	IV
863	0	0	0	0	0	0	0	0	II	0
866	0	0	0	0	0	IV	0	0	0	0
867	0	0	0	0	0	0	II	0	0	0
869	0	0	0	0	0	0	0	0	0	0
872	I	I	IV	IV	I	IV	I	IV	III	II
873	0	II	III	0	III	0	IV	II	II	0
878	II	II	0	0	II	0	0	0	I	II
882	II	II	III	I	0	II	IV	I	I	II
883	IV	IV	II	III	IV	II	IV	III	IV	IV
886	I	IV	I	I	I	I	IV	I	II	II
887	I	0	0	0	0	0	0	0	0	0
892	0	I	II	0	0	0	0	III	I	I
901	IV	II	III	II	IV	0	IV	I	IV	I
913	0	0	0	II	0	0	I	0	0	0
927	I	I	II	IV	I	IV	I	I	I	II
931	0	0	0	II	0	0	0	0	I	IV
932	I	0	0	0	0	0	0	II	0	0
933	0	0	0	0	0	0	0	0	0	0
937	III	II	IV	III	III	II	IV	II	III	I
938	0	0	II	0	0	II	II	0	0	0
939	III	0	0	0	0	0	0	0	0	0
940	0	0	0	0	II	0	0	0	0	II
943	0	0	0	0	III	0	0	0	IV	0
951	0	II	0	0	0	0	0	0	0	0
955	0	I	0	0	I	0	II	0	I	II
962	IV	I	I	I	IV	I	I	I	IV	I
966	I	0	I	III	I	0	0	0	0	I
973	0	II	0	II	0	0	0	III	0	0
975	0	0	0	0	0	0	0	0	0	0
978	0	0	0	0	0	III	0	0	0	0
983	I	II	IV	II	I	II	I	I	II	I
995	0	0	II	0	0	II	0	III	0	III
998	0	0	0	0	0	II	II	0	0	0
1007	0	0	0	0	0	0	0	0	0	0
1014	I	I	I	I	II	I	IV	I	I	I
1019	II	0	II	I	0	0	II	0	0	0
1020	II	II	0	I	I	0	I	0	II	0
1021	II	II	0	0	II	0	0	0	0	0
1025	0	0	II	0	0	0	0	IV	0	0
1028	0	0	0	0	0	0	0	0	0	0
1032	0	0	0	0	0	0	0	0	0	0
1034	II	0	0	IV	0	0	0	0	0	II

No. Paciente	1	2	3	4	5	6	7	8	9	10
1035	0	0	0	0	0	0	0	0	0	II
1042	0	0	0	0	0	0	0	0	0	0
1046	0	II	II	0	0	0	0	0	0	II
1056	0	0	0	0	0	0	0	0	0	0
1059	0	IV	I	I	I	I	0	0	IV	0
1061	0	0	0	0	0	0	0	0	II	0
1067	III	III	III	III	III	IV	III	III	II	II
1070	0	0	0	0	0	0	0	0	0	0
1080	0	0	0	0	II	0	0	0	0	II
1081	0	0	0	0	0	0	0	0	0	0
1085	IV	0	0	0	0	0	0	II	0	0
1087	0	0	0	0	0	0	0	II	II	0
1091	IV	0	II	0	0	0	0	0	0	0
1099	II	I	I	III	I	II	I	II	III	III
1107	I	II	I	I	I	II	II	I	I	I
1108	0	0	0	0	II	II	II	II	0	0
1110	0	0	0	0	0	0	0	0	0	0
1115	0	0	0	II	0	0	0	0	0	0
1122	III	IV	II	IV	III	IV	II	IV	I	I
1123	IV	IV	IV	IV	II	IV	IV	IV	IV	IV
1125	II	II	IV	IV	II	III	I	III	IV	IV
1126	I	I	I	I	II	I	I	I	I	II
1133	I	II	I	II	I	II	I	I	II	0
1136	I	IV	II	III	II	II	II	IV	I	II
1137	0	0	0	0	0	0	0	0	0	0
1139	I	I	I	II	IV	II	I	I	I	III
1140	0	0	0	0	II	II	0	0	0	0
1146	0	0	0	0	0	0	0	0	0	0
1155	0	0	II	0	0	0	0	0	0	0
1156	0	0	0	0	0	0	0	0	0	IV
1176	I	0	0	0	0	II	I	I	0	IV
1177	II	I	I	I	II	I	II	I	I	I
1181	0	0	0	0	0	0	IV	0	II	0
1184	0	II	0	0	0	I	I	II	I	0
1185	0	0	0	0	0	0	0	0	0	0
1186	0	0	0	II	0	0	0	0	0	0
1187	0	0	0	0	0	0	0	0	0	0
1188	II	I	I	I	I	II	IV	II	I	I
1192	0	0	0	0	0	0	0	0	0	0
1195	0	0	0	0	0	0	0	0	0	0
1196	0	III	0	0	0	II	0	0	0	0
1198	IV	II	III	II	IV	II	II	II	II	I
1202	0	0	0	0	0	0	II	0	0	0
1203	I	I	0	0	0	III	0	0	0	I
1204	0	0	0	0	0	0	0	0	0	0
1205	0	0	0	0	0	0	0	0	0	0
1206	0	I	II	0	I	I	0	II	I	I
1207	II	0	0	0	0	0	0	0	0	II
1210	II	II	III	II	III	II	IV	I	III	II
1212	I	I	II	I	I	I	IV	I	I	II
1215	0	0	0	0	0	II	0	II	0	0
1216	I	I	I	I	II	II	I	I	I	I
1217	0	0	0	0	0	0	0	0	III	0
1220	II	II	II	II	0	0	III	0	0	III
1222	0	0	0	0	0	0	II	III	II	0
1223	0	0	0	0	0	0	III	0	0	0
1224	0	II	II	I	0	I	I	0	0	I
1230	I	II	I	0	I	I	II	II	0	I
1231	0	II	0	0	II	0	0	0	0	0
1236	0	II	0	0	II	0	0	III	0	0
1240	0	0	0	0	0	II	0	II	0	0
1247	0	II	0	0	0	0	0	II	0	II
1250	0	0	0	0	0	0	0	0	0	0
1257	0	II	II	0	0	0	0	0	0	0
1259	0	0	0	0	0	II	0	0	0	0
1263	III	III	III	III	III	III	III	III	III	III
1264	0	0	0	0	II	IV	0	II	0	0
1267	0	0	0	0	0	0	II	I	0	0
1273	0	0	0	0	0	0	0	0	0	II
1274	0	0	0	0	0	0	0	II	II	0
1277	0	0	II	0	0	0	0	0	0	0
1281	0	II	0	0	II	II	0	0	0	0
1282	I	0	0	0	I	0	0	0	0	0
1292	I	I	I	III	I	I	I	II	III	I
1293	0	0	0	0	0	0	0	0	0	0
1302	0	II	0	0	0	0	II	II	0	II
1308	II	0	0	0	0	IV	0	0	0	IV
1311	0	0	0	0	0	0	0	0	0	0
1316	II	I	I	II	0	III	I	III	I	I
1317	I	III	III	II	III	III	II	II	III	I

Imputaciones de la variable ***etapa*** (*Continuación*)

No. Paciente	1	2	3	4	5	6	7	8	9	10
1318	IV	I	II	IV	III	IV	II	0	IV	II
1326	II	0	0	0	0	0	0	0	0	0
1327	0	0	0	0	0	0	0	II	0	0
1334	II	0	0	0	II	0	0	0	0	0
1336	0	0	0	0	0	0	IV	0	0	II
1337	0	0	I	0	0	I	I	0	II	0
1347	I	I	II	I	0	I	0	IV	I	IV
1351	II	I	II	II	I	II	IV	III	I	I
1355	II	II	IV	0	I	IV	0	I	I	II
1357	II	III	II	I	I	III	III	I	IV	IV
1359	I	I	I	II	I	I	I	I	I	I
1360	0	0	II	II	0	0	0	0	0	0
1362	0	0	0	0	0	0	0	0	0	0
1364	0	0	0	0	0	0	0	0	0	0
1366	0	0	0	0	0	II	IV	0	IV	0
1368	I	II	II	II	II	II	II	I	II	II
1372	0	0	0	II	0	0	0	II	II	0
1375	0	0	0	II	0	0	0	0	0	0
1376	I	I	I	0	0	I	0	IV	I	III
1383	0	0	0	III	0	0	0	0	IV	0
1384	0	0	0	0	0	0	0	0	0	0
1398	0	0	0	0	0	II	0	0	0	0
1404	I	I	I	I	I	I	I	I	I	I
1405	0	0	0	0	0	0	0	II	0	0
1406	0	II	0	0	I	II	0	0	0	I
1411	0	0	0	0	0	0	0	0	0	II
1412	0	0	0	0	0	III	0	0	0	0
1416	0	0	II	II	0	0	II	IV	0	0
1428	I	I	I	I	I	I	II	I	I	II
1435	0	0	0	0	0	III	II	0	0	0
1436	0	0	0	0	0	0	II	0	0	0
1440	0	0	II	II	0	0	0	0	0	0
1442	0	0	0	0	0	0	0	0	0	0
1443	0	0	0	0	0	II	0	0	0	0
1448	0	0	0	0	0	0	0	II	0	0
1453	0	II	0	IV	II	II	0	0	II	0
1458	II	II	III	II	III	II	II	III	III	II
1464	0	0	0	I	0	0	0	0	I	0
1474	II	0	0	0	0	0	II	0	0	0
1475	0	II	0	0	0	0	0	0	II	0
1477	II	0	0	II	0	0	0	II	0	0
1478	IV	II	II	IV	IV	III	IV	IV	IV	IV
1491	0	0	0	II	0	0	II	II	0	II
1500	IV	I	II	II	II	II	I	I	II	I
1516	II	0	0	0	0	0	0	II	0	0
1518	0	II	0	0	0	II	0	0	0	0
1523	0	0	0	0	0	0	0	0	0	II
1525	0	0	0	0	0	II	0	0	0	0
1529	II	0	II	I	II	0	0	0	0	0
1533	I	0	0	0	0	0	0	0	II	0
1535	0	0	0	0	0	0	II	II	II	0
1539	0	0	0	0	II	0	0	II	0	0
1543	0	0	0	II	0	0	II	0	0	0
1544	II	0	0	0	II	0	0	0	0	0
1546	0	0	0	0	0	0	0	II	II	II
1547	I	I	I	I	I	III	II	I	I	I
1550	0	II	0	0	0	0	0	0	0	0
1560	I	II	I	I	II	I	I	I	I	I
1567	0	0	0	0	0	II	0	0	0	0
1581	0	0	0	0	0	0	0	0	0	II
1589	0	0	0	0	0	0	0	0	0	0
1593	I	II	II	II	I	I	I	I	III	II
1600	0	0	0	0	0	0	0	0	0	0
1603	0	0	0	0	0	0	0	0	0	0
1605	0	0	0	0	II	0	0	0	0	0
1610	II	III	I	II	III	I	I	III	II	II
1613	0	II	0	0	0	0	0	0	0	0
1620	I	I	I	I	I	I	I	I	I	I
1622	IV	II	IV	I	III	I	I	III	IV	I
1624	0	0	II	II	0	0	0	0	0	0
1628	III	II	I	III	I	II	I	II	I	II
1629	II	I	I	I	I	II	I	I	I	I
1630	0	0	0	0	0	0	0	0	0	0
1637	II	0	0	0	II	0	0	II	0	0
1640	0	0	0	0	II	0	0	0	0	0
1642	II	II	I	IV	IV	II	II	I	I	II
1654	0	0	0	0	II	0	0	II	0	0
1655	0	IV	II	0	0	0	0	0	0	II
1660	II	I	II	III	IV	III	IV	I	II	IV
1673	0	0	0	0	0	0	0	0	0	0

No. Paciente	1	2	3	4	5	6	7	8	9	10
1674	I	0	0	I	0	II	0	0	0	0
1679	II	0	0	0	0	0	0	0	0	II
1684	0	0	0	0	0	0	0	0	0	0
1686	0	0	0	0	0	0	0	0	0	0
1691	I	II	I	II	I	I	I	I	I	I
1692	III	I	I	I	I	I	I	II	I	I
1701	0	0	II	II	II	0	0	0	0	0
1702	0	I	I	I	I	0	I	I	0	I
1706	0	0	0	0	0	0	0	0	0	0
1713	I	0	0	0	0	II	0	II	I	I
1715	0	0	II	0	0	II	0	0	0	0
1718	0	0	II	0	0	II	0	0	0	0
1720	II	0	0	0	0	0	0	0	0	0
1724	0	0	0	II	0	0	0	0	0	0
1727	II	0	0	II	0	0	0	II	II	II
1728	0	0	II	0	0	0	0	0	II	0
1730	0	II	II	0	0	0	0	0	0	0
1736	II	I	I	III	IV	III	III	I	II	I
1737	IV	I	IV	II	IV	IV	II	IV	IV	II
1748	0	0	II	0	0	0	0	0	0	II
1758	I	I	I	I	0	I	0	0	0	I
1759	0	0	0	0	0	II	0	II	II	0
1760	0	0	0	0	0	0	0	0	0	0
1769	0	0	II	0	0	0	IV	II	0	0
1770	II	0	0	II	0	0	0	0	I	0
1771	II	0	0	II	0	0	0	0	0	II
1773	I	0	0	I	I	I	I	0	0	0
1774	0	II	0	0	0	0	II	II	0	II
1787	0	0	I	II	I	I	0	0	0	0
1789	0	0	0	0	0	0	0	II	0	0
1790	0	0	0	0	II	0	0	0	0	0
1796	II	0	I	0	I	I	0	0	I	0
1800	I	I	II	IV	I	0	I	II	0	0
1801	0	0	0	0	0	0	0	0	0	II
1802	0	0	0	II	0	0	0	0	II	0
1804	0	0	0	0	0	0	0	0	0	0
1810	0	0	0	0	0	0	0	0	0	0
1822	0	0	II	0	0	0	0	0	0	II
1825	0	I	I	0	I	I	I	0	I	0
1827	0	0	II	0	0	II	0	0	0	0
1832	0	0	0	0	0	0	0	0	0	0
1836	IV	II	II	II	II	0	II	0	III	II
1846	I	I	I	II	I	I	I	I	I	I
1850	0	II	0	0	0	0	II	0	0	0
1853	0	II	II	0	0	0	0	II	0	0
1854	0	0	0	0	0	0	0	0	0	II
1860	0	0	0	0	0	0	0	II	IV	0
1868	II	0	0	0	II	IV	0	0	0	II
1869	IV	IV	II	II	IV	IV	IV	II	II	I
1874	0	I	I	I	0	0	0	0	0	I
1875	0	0	0	I	0	0	I	0	0	II
1877	0	0	0	I	0	0	0	0	0	0
1878	IV	0	0	0	0	0	0	0	0	0
1881	0	0	0	0	0	0	0	0	0	0
1882	0	0	0	0	0	0	0	II	0	0
1886	III	III	III	III	IV	III	IV	III	III	II
1890	II	III	IV	IV	II	I	IV	II	II	II
1892	0	0	II	0	II	0	0	0	0	0
1900	I	II	I	I	I	II	II	II	II	II
1903	III	III	II	IV	IV	III	II	III	III	II
1905	0	0	0	0	0	II	0	0	I	II
1921	II	0	0	0	II	0	0	0	II	0
1925	IV	0	0	II	II	II	II	II	II	III
1928	I	I	I	I	I	I	I	I	I	I
1933	IV	0	III	II	0	0	0	II	0	0
1935	II	II	II	I	III	I	II	I	II	I
1938	0	0	II	II	0	I	I	I	I	I
1940	II	0	0	0	0	0	0	0	0	0
1942	0	0	II	II	0	0	0	0	0	0
1950	IV	I	II	I	I	I	I	I	I	II
1953	0	0	II	II	0	0	I	0	0	0
1965	IV	I	I	0	0	I	0	0	0	0
1969	I	I	II	III	III	II	I	I	I	I
1974	0	II	0	0	0	I	II	II	0	0
1975	0	0	0	0	0	III	0	II	0	0
1976	II	II	0	III	III	II	II	III	III	II
1979	0	0	0	II	0	0	0	II	0	0
1980	I	0	II	I	0	II	I	0	0	0
1986	III	II	III	II	III	III	III	I	III	I
1987	I	II	I	II	IV	I	IV	IV	I	I

Imputaciones de la variable ***etapa*** (*Continuación*)

No. Paciente	1	2	3	4	5	6	7	8	9	10
1996	0	0	0	0	0	II	0	0	II	0
1997	II	II	II	0	0	0	0	I	0	0
2010	0	0	0	II	0	0	0	0	II	0
2011	0	0	0	0	0	0	0	0	0	0
2020	0	0	0	0	0	II	0	0	0	0
2027	I	I	I	I	I	II	I	I	II	I
2036	III	III	III	IV	III	III	III	II	III	III
2037	II	I	III	II	I	IV	II	I	IV	II
2039	0	0	0	II	0	0	0	0	0	0
2054	0	0	0	0	0	0	0	II	II	0
2056	0	0	0	0	0	0	0	0	0	0
2063	II	III	I	IV	II	II	III	III	I	I
2069	0	0	II	0	0	0	0	II	0	0
2075	0	0	0	0	0	0	0	0	II	II
2076	0	0	0	0	III	0	II	0	II	II
2081	0	0	0	II	0	0	0	0	0	II
2082	0	0	II	II	0	II	0	0	0	0
2091	III	I	I	III	I	I	I	I	III	III
2093	0	II	II	IV	0	0	II	0	0	0
2096	0	0	II	0	II	II	III	0	II	0
2100	0	0	0	II	0	0	0	0	0	0
2110	II	I	II	I	I	II	I	II	I	I
2112	IV	III	II	I	II	I	II	I	III	II
2114	0	0	0	II	0	0	0	0	0	0
2115	II	IV	0	0	0	0	0	0	0	0
2117	0	0	0	0	0	III	II	0	0	0
2118	0	III	0	0	0	0	0	0	0	0
2128	0	I	0	0	I	I	I	II	III	0
2130	II	0	II	0	0	II	0	0	III	0
2144	0	0	0	0	0	0	0	II	0	0
2154	0	0	0	0	0	0	II	0	0	0
2157	0	0	0	0	IV	0	0	0	0	II
2166	0	0	0	II	IV	0	0	II	0	0
2168	0	0	0	0	II	0	0	0	0	II
2169	0	0	0	0	0	0	0	0	0	II
2171	0	0	0	0	II	I	0	I	0	II
2176	0	0	I	I	0	0	I	0	I	0
2179	I	I	I	II	I	I	I	II	I	I
2186	I	I	III	I	I	I	I	II	I	0
2189	0	0	0	0	0	0	0	0	0	0
2190	II	0	0	0	0	0	II	0	II	0
2191	0	0	0	0	0	II	0	II	0	0
2196	0	0	0	0	0	0	0	0	0	0
2200	III	II	I	I	I	I	I	II	III	II
2202	0	0	0	0	0	II	0	0	0	II
2210	0	0	0	0	0	II	III	0	II	0
2212	I	II	II	IV	II	II	II	II	III	IV
2217	0	0	0	0	II	0	0	0	0	II
2221	I	II	IV	II	II	II	II	II	I	II
2222	II	I	IV	II	I	III	II	IV	II	IV
2225	0	0	II	0	0	0	0	0	0	0
2236	0	II	III	0	0	0	0	III	II	0
2238	II	II	0	II	III	II	II	II	II	II
2239	0	0	0	0	0	0	0	0	0	II
2241	II	0	0	0	0	II	II	0	0	II
2245	II	I	III	III	IV	III	III	III	III	II
2247	0	II	0	0	0	II	0	0	0	II
2267	0	II	II	0	III	0	III	0	0	0
2273	0	0	0	II	0	0	0	0	0	II
2274	I	II	I	I	I	I	I	I	I	II
2279	III	III	IV	III	III	III	III	II	III	III
2281	0	II	0	0	0	0	0	II	0	II
2286	0	0	0	0	0	0	II	0	0	0
2288	0	II	0	0	0	0	0	0	0	0
2292	0	III	II	II	III	II	I	I	II	III
2295	0	0	0	0	II	0	0	0	0	0
2309	0	0	0	II	II	0	0	0	II	0
2314	0	I	I	I	I	II	0	I	0	I
2319	IV	II	II	III	III	III	III	III	0	IV
2322	0	II	0	0	0	0	0	0	0	II
2323	0	0	0	II	0	0	0	0	II	0
2326	II	0	0	IV	0	0	0	0	0	0
2333	0	0	0	II	0	0	0	0	0	0
2335	0	II	0	0	0	0	0	0	0	0
2337	I	I	0	I	II	I	II	I	II	I
2341	I	I	I	I	II	I	II	I	II	II
2344	0	0	0	0	0	0	0	0	0	0
2348	0	0	0	0	II	0	0	0	0	II
2350	0	0	0	0	0	II	0	II	0	0
2354	0	0	0	0	0	0	II	0	0	0
2355	0	II	0	I	0	II	0	I	I	0

No. Paciente	1	2	3	4	5	6	7	8	9	10
2358	III	III	I	III	II	III	I	III	I	I
2362	III	II	III	II	IV	II	III	II	III	II
2365	0	0	0	II	0	0	II	0	0	0
2369	0	0	0	0	0	0	0	0	0	II
2373	0	0	0	0	0	0	0	II	0	0
2380	0	0	0	II	0	0	0	0	0	0
2388	0	0	0	0	0	II	0	II	0	0
2401	0	0	0	0	0	0	0	0	0	0
2402	I	I	I	I	I	I	I	I	0	I
2406	0	0	0	0	0	0	0	0	0	II
2420	0	0	II	0	0	0	0	II	II	0
2423	0	II	0	0	0	0	II	0	0	0
2432	II	II	II	0	0	II	II	III	0	0
2434	II	I	I	0	0	0	0	0	I	II
2438	0	0	0	0	0	0	II	0	0	0
2443	0	0	0	0	0	0	0	0	0	0
2444	I	I	II	III	I	I	II	I	I	I
2451	I	I	I	I	I	I	I	I	I	I
2452	0	0	I	I	0	0	II	I	0	II
2456	II	II	0	0	0	0	I	0	0	I
2467	0	0	0	0	0	0	0	0	0	0
2468	0	0	0	II	0	II	0	0	II	0
2477	0	II	0	II	0	0	0	0	0	0
2482	0	0	0	0	II	0	0	0	II	0
2483	0	0	II	0	0	0	0	0	0	0
2488	II	III	0	0	0	0	II	0	0	II
2495	0	0	0	0	0	0	0	0	0	0
2496	0	0	0	0	0	0	0	0	0	0
2497	I	I	I	I	I	I	I	I	I	I
2502	0	0	0	0	0	0	II	0	0	II
2503	II	II	II	III	II	III	III	III	III	II
2510	II	I	I	I	II	II	I	I	I	I
2512	I	I	0	0	0	I	0	0	II	0
2513	II	0	0	0	0	0	0	0	0	0
2519	I	I	I	I	I	I	I	II	III	III
2521	0	0	II	0	0	0	0	0	III	0
2523	0	0	0	0	III	0	IV	0	0	0
2527	0	0	II	0	0	0	0	0	0	II
2531	0	0	0	0	0	0	0	0	0	0
2532	I	0	II	I	0	I	II	0	0	0
2534	0	II	II	0	0	II	0	0	0	II
2539	0	0	0	0	0	0	0	0	0	0
2542	0	0	0	0	0	0	0	0	0	0
2548	0	II	0	0	0	0	0	0	0	0
2549	0	0	0	0	0	0	I	0	0	0
2550	II	0	0	0	0	0	0	0	II	II
2552	0	0	0	0	0	0	0	0	0	II
2553	0	0	II	0	0	0	0	0	0	0
2554	I	I	I	I	I	I	I	I	I	I
2559	0	0	I	0	II	0	I	0	0	0
2560	0	0	0	0	0	0	0	0	0	0
2561	0	0	0	0	II	0	0	0	0	0
2571	0	0	0	II	0	0	0	0	0	0
2574	0	0	0	0	0	0	0	0	II	0
2576	0	0	II	0	0	0	0	0	0	0
2579	III	IV	II	II	II	II	III	IV	II	II
2580	0	0	0	0	0	0	0	0	0	0
2581	0	0	0	0	0	0	0	II	0	0
2588	III	III	IV	I	III	III	IV	I	III	III
2595	I	III	I	III	II	I	0	II	III	II
2597	II	0	II	0	0	0	0	0	0	0
2598	0	0	0	0	II	II	0	0	0	0
2601	II	III	0	II	0	II	0	III	II	II
2605	0	0	0	0	0	0	0	0	0	0
2614	0	0	0	0	0	0	0	0	II	0
2615	II	0	0	0	0	0	0	0	0	0
2619	0	0	0	0	0	0	0	II	0	II
2620	0	0	0	0	0	II	II	0	0	0
2621	I	IV	I	III	II	I	II	I	II	I
2623	0	0	0	0	0	0	II	0	0	0
2630	I	II	III	I	II	II	II	III	I	I
2634	0	0	0	II	0	0	0	0	II	0
2636	0	0	0	0	0	0	0	0	0	0
2640	0	0	0	0	0	0	0	0	0	0
2642	I	III	I	IV	0	I	II	I	II	I
2644	0	0	II	0	0	0	0	II	II	II
2656	II	II	0	0	0	II	0	0	0	0
2666	0	0	0	0	0	0	0	0	0	0
2671	0	II	II	0	0	0	0	0	0	0
2680	II	II	I	II	II	II	I	I	II	II
2681	II	0	0	0	II	II	0	II	I	0
2684	0	0	0	0	0	0	0	0	0	0

Apéndice C

Estimación de los parámetros de la función de Larson & Dinse por imputación

IMPUTACIÓN 1

Estimación de los coeficientes del modelo

Variables	α	β_1	β_2
Intercepción	-1.27367	–	–
	(0.443263325)	–	–
edad	-0.01223	0.05324	0.07538
	(0.000077887)	(0.000096586)	(0.000013050)
etapa I	0.95571	0.12841	-0.00152
	(0.029259513)	(0.037879071)	(0.004539531)
etapa II	0.86721	0.08685	-0.10064
	(0.033996257)	(0.046037678)	(0.005073107)
etapa III	1.33301	0.78653	-0.07343
	(0.055461398)	(0.062081366)	(0.011918975)
etapa IV	2.68034	1.17334	0.34841
	(0.045785102)	(0.039943383)	(0.025813981)
civil otro	0.21441	0.45033	0.27213
	(0.018900049)	(0.017568906)	(0.003706495)
tratamiento II	-0.20673	-0.00234	-0.03766
	(0.024736077)	(0.024342303)	(0.005303862)
tratamiento III	-0.36782	-0.79123	-0.34531
	(0.043988202)	(0.051322648)	(0.008628381)

Parámetros de la función de supervivencia condicional base

$\log(\lambda_1)$	$\log(\lambda_2)$	$\log(\gamma_1)$	$\log(\gamma_2)$
5.09310	6.45274	0.38569	0.31086
(0.273643502)	(0.039562799)	(0.001964105)	(0.039562799)

IMPUTACIÓN 2

Estimación de los coeficientes del modelo

Variables	α	β_1	β_2
Intercepción	-1.32168	–	–
	(0.422950769)	–	–
edad	-0.01134	0.05395	0.07568
	(0.000074653)	(0.000093434)	(0.000012805)
etapa I	0.92233	0.24288	-0.01671
	(0.029250254)	(0.036367621)	(0.004434308)
etapa II	0.91498	-0.02951	-0.04667
	(0.034483029)	(0.045860697)	(0.005139936)
etapa III	1.50280	0.54777	-0.06652
	(0.053153138)	(0.056171094)	(0.011965452)
etapa IV	2.73398	1.29636	0.38942
	(0.044869439)	(0.037361571)	(0.023826050)
civil otro	0.17705	0.44506	0.27199
	(0.018828254)	(0.016800391)	(0.003659269)
tratamiento II	-0.23027	-0.06622	-0.03573
	(0.024241320)	(0.022621843)	(0.005191888)
tratamiento III	-0.47948	-0.69411	-0.35931
	(0.038776569)	(0.042285225)	(0.007976033)

Parámetros de la función de supervivencia condicional base

$\log(\lambda_1)$	$\log(\lambda_2)$	$\log(\gamma_1)$	$\log(\gamma_2)$
5.06642	6.47776	0.40000	0.31004
(0.256001886)	(0.038763957)	(0.001882679)	(0.038763957)

IMPUTACIÓN 3

Estimación de los coeficientes del modelo

Variables	α	β_1	β_2
Intercepción	-1.15657	–	–
	(0.439513465)	–	–
edad	-0.01477	0.05622	0.07538
	(0.000077548)	(0.000093057)	(0.000013130)
etapa I	0.96605	0.10234	-0.00320
	(0.030329864)	(0.038705648)	(0.004391338)
etapa II	1.06475	-0.06784	-0.09004
	(0.033435868)	(0.044933059)	(0.005199401)
etapa III	1.60119	0.59411	0.00934
	(0.056006962)	(0.060584958)	(0.012336896)
etapa IV	2.95472	0.92034	0.56719
	(0.053465820)	(0.039838662)	(0.047797880)
civil otro	0.21215	0.44757	0.27714
	(0.019281963)	(0.017191789)	(0.003720689)
tratamiento II	-0.20578	-0.13106	-0.02366
	(0.024811979)	(0.024216291)	(0.005165303)
tratamiento III	-0.41197	-0.88575	-0.34847
	(0.045626933)	(0.048419569)	(0.008836454)

Parámetros de la función de supervivencia condicional base

$\log(\lambda_1)$	$\log(\lambda_2)$	$\log(\gamma_1)$	$\log(\gamma_2)$
5.22053	6.45246	0.35499	0.31362
(0.286457067)	(0.039436781)	(0.001984379)	(0.039436781)

IMPUTACIÓN 4

Estimación de los coeficientes del modelo

Variables	α	β_1	β_2
Intercepción	-1.24628	–	–
	(0.420899868)	–	–
edad	-0.01206	0.04941	0.07551
	(0.000074240)	(0.000092244)	(0.000012905)
etapa I	0.82118	0.17052	-0.00660
	(0.029144769)	(0.037448957)	(0.004372357)
etapa II	0.89115	-0.12893	-0.02835
	(0.033652917)	(0.045938104)	(0.005193776)
etapa III	1.35812	0.67510	-0.01194
	(0.052752275)	(0.058829860)	(0.011579427)
etapa IV	2.76610	0.91844	0.49148
	(0.047436194)	(0.036135211)	(0.032275610)
civil otro	0.15776	0.44774	0.27246
	(0.019114415)	(0.017900733)	(0.003692082)
tratamiento II	-0.21148	-0.12793	-0.01915
	(0.024351359)	(0.023360005)	(0.005174809)
tratamiento III	-0.37176	-0.86628	-0.34964
	(0.044550978)	(0.049987397)	(0.008717497)

Parámetros de la función de supervivencia condicional base

$\log(\lambda_1)$	$\log(\lambda_2)$	$\log(\gamma_1)$	$\log(\gamma_2)$
4.85985	6.47591	0.36183	0.31220
(0.275800956)	(0.038825193)	(0.001905922)	(0.038825193)

IMPUTACIÓN 5

Estimación de los coeficientes del modelo

Variables	α	β_1	β_2
Intercepción	-1.28027	–	–
	(0.430139239)	–	–
edad	-0.01102	0.05960	0.07556
	(0.000075082)	(0.000089440)	(0.000012975)
etapa I	0.88811	0.26932	0.03261
	(0.028725570)	(0.035363138)	(0.004478717)
etapa II	0.86723	0.00492	-0.02499
	(0.033595945)	(0.041670778)	(0.005196820)
etapa III	1.34074	0.75954	0.10742
	(0.053377133)	(0.057987775)	(0.011406312)
etapa IV	2.68819	1.23541	0.51701
	(0.044865684)	(0.035761142)	(0.025331885)
civil otro	0.20766	0.41959	0.28485
	(0.018787978)	(0.017091028)	(0.003682885)
tratamiento II	-0.26143	-0.00364	-0.02030
	(0.024225863)	(0.021833333)	(0.005223302)
tratamiento III	-0.46063	-0.69224	-0.40624
	(0.039622451)	(0.045429097)	(0.008184896)

Parámetros de la función de supervivencia condicional base

$\log(\lambda_1)$	$\log(\lambda_2)$	$\log(\gamma_1)$	$\log(\gamma_2)$
5.40074	6.49472	0.39368	0.31052
(0.252206899)	(0.039466433)	(0.001941024)	(0.039466433)

IMPUTACIÓN 6

Estimación de los coeficientes del modelo

Variables	α	β_1	β_2
Intercepción	-1.24317	–	–
	(0.444364893)	–	–
edad	-0.01266	0.06189	0.07541
	(0.000077072)	(0.000092000)	(0.000013057)
etapa I	0.97931	0.21176	0.01896
	(0.029920446)	(0.037012282)	(0.004519986)
etapa II	0.89615	0.04969	-0.03866
	(0.034365122)	(0.045851268)	(0.005049395)
etapa III	1.52296	0.78616	0.02259
	(0.053058855)	(0.056174104)	(0.011497260)
etapa IV	2.75445	1.28179	0.41640
	(0.046873470)	(0.040328347)	(0.028019469)
civil otro	0.22269	0.43102	0.28301
	(0.019028169)	(0.017129993)	(0.003706893)
tratamiento II	-0.23551	0.04357	-0.04026
	(0.024578505)	(0.022839259)	(0.005300194)
tratamiento III	-0.47798	-0.64996	-0.39307
	(0.041700681)	(0.047726942)	(0.008382746)

Parámetros de la función de supervivencia condicional base

$\log(\lambda_1)$	$\log(\lambda_2)$	$\log(\gamma_1)$	$\log(\gamma_2)$
5.55739	6.46716	0.39309	0.31033
(0.264078301)	(0.040087896)	(0.002004229)	(0.040087896)

IMPUTACIÓN 7

Estimación de los coeficientes del modelo

Variables	α	β_1	β_2
Intercepción	-1.28661	–	–
	(0.427273668)	–	–
edad	-0.01140	0.05449	0.07540
	(0.000074614)	(0.000098074)	(0.000012924)
etapa I	0.84083	0.24182	0.01172
	(0.029330057)	(0.036602459)	(0.004404849)
etapa II	0.87695	0.04098	-0.06208
	(0.033630834)	(0.045134602)	(0.005149197)
etapa III	1.35314	0.77778	-0.04165
	(0.054771521)	(0.061806934)	(0.011862505)
etapa IV	2.66166	1.17479	0.56400
	(0.046288963)	(0.040702104)	(0.033313686)
civil otro	0.19517	0.40309	0.27729
	(0.018837310)	(0.017538144)	(0.003696623)
tratamiento II	-0.21658	-0.06195	-0.01595
	(0.024314969)	(0.023373479)	(0.005263136)
tratamiento III	-0.43303	-0.81243	-0.35551
	(0.046114006)	(0.057757097)	(0.008834017)

Parámetros de la función de supervivencia condicional base

$\log(\lambda_1)$	$\log(\lambda_2)$	$\log(\gamma_1)$	$\log(\gamma_2)$
5.13533	6.46950	0.38526	0.31278
(0.283010531)	(0.039057474)	(0.002129563)	(0.039057474)

IMPUTACIÓN 8

Estimación de los coeficientes del modelo

Variables	α	β_1	β_2
Intercepción	-1.24597	–	–
	(0.436271525)	–	–
edad	-0.01280	0.06040	0.07533
	(0.000076354)	(0.000094080)	(0.000013155)
etapa I	0.95566	0.26256	-0.00424
	(0.029399151)	(0.037672211)	(0.004464943)
etapa II	0.92028	-0.06751	-0.04300
	(0.034762136)	(0.047046323)	(0.005195260)
etapa III	1.30230	0.83991	-0.03276
	(0.055248636)	(0.063925452)	(0.011513263)
etapa IV	2.87349	1.20083	0.73515
	(0.053822023)	(0.043198928)	(0.056939301)
civil otro	0.20558	0.47329	0.27618
	(0.019282569)	(0.018771825)	(0.003785498)
tratamiento II	-0.20814	-0.01233	-0.02613
	(0.024887246)	(0.023474844)	(0.005242406)
tratamiento III	-0.28968	-0.83217	-0.35876
	(0.051772270)	(0.065953930)	(0.009711633)

Parámetros de la función de supervivencia condicional base

$\log(\lambda_1)$	$\log(\lambda_2)$	$\log(\gamma_1)$	$\log(\gamma_2)$
5.47484	6.45102	0.38306	0.31418
(0.263001710)	(0.039619695)	(0.002118180)	(0.039619695)

IMPUTACIÓN 9

Estimación de los coeficientes del modelo

Variables	α	β_1	β_2
Intercepción	-1.09302	–	–
	(0.441770856)	–	–
edad	-0.01470	0.06277	0.07505
	(0.000077998)	(0.000092138)	(0.000013098)
etapa I	0.96496	0.07443	0.02048
	(0.030226699)	(0.037943804)	(0.004490183)
etapa II	1.00115	-0.15115	-0.02969
	(0.034407788)	(0.044363704)	(0.005268065)
etapa III	1.43461	0.60423	-0.01939
	(0.053875693)	(0.055528782)	(0.011155117)
etapa IV	2.84356	1.21845	0.50534
	(0.046159128)	(0.037641843)	(0.026915005)
civil otro	0.13345	0.46272	0.27333
	(0.019283615)	(0.017063765)	(0.003676501)
tratamiento II	-0.22995	-0.07391	-0.03010
	(0.024797596)	(0.022675277)	(0.005204909)
tratamiento III	-0.47556	-0.68273	-0.38181
	(0.040972046)	(0.045323010)	(0.008266795)

Parámetros de la función de supervivencia condicional base

$\log(\lambda_1)$	$\log(\lambda_2)$	$\log(\gamma_1)$	$\log(\gamma_2)$
5.45768	6.44723	0.40308	0.31145
(0.254682033)	(0.039809086)	(0.001983600)	(0.039809086)

IMPUTACIÓN 10

Estimación de los coeficientes del modelo

Variables	α	β_1	β_2
Intercepción	-1.20071	–	–
	(0.442883652)	–	–
edad	-0.01201	0.05508	0.07541
	(0.000077013)	(0.000091551)	(0.000013113)
etapa I	0.84970	0.26510	0.03381
	(0.029559631)	(0.037892199)	(0.004502901)
etapa II	0.74911	0.11416	-0.04698
	(0.032866644)	(0.042819425)	(0.004786170)
etapa III	1.29489	0.80698	0.04401
	(0.054894328)	(0.060360154)	(0.012254927)
etapa IV	2.82673	1.14281	0.70214
	(0.050403152)	(0.038306656)	(0.044437276)
civil otro	0.18876	0.47807	0.27631
	(0.019194081)	(0.017775305)	(0.003731849)
tratamiento II	-0.25352	-0.01310	-0.02041
	(0.024736939)	(0.022996674)	(0.005279396)
tratamiento III	-0.31299	-0.84787	-0.36408
	(0.044258568)	(0.048731335)	(0.008912644)

Parámetros de la función de supervivencia condicional base

$\log(\lambda_1)$	$\log(\lambda_2)$	$\log(\gamma_1)$	$\log(\gamma_2)$
5.27415	6.46698	0.36782	0.31404
(0.279156718)	(0.039969035)	(0.001984386)	(0.039969035)

Bibliografía

[1] Larson M.G. y Dinse G.E. (1985). A mixture model for the regression analysis of competing risks data. Appl Statist, 34: 201-211.

[2] Maller R.A. y Zhou X. (2002). Analysis of Parametric Models for Competing Risks. Statistica Sinica, 12: 725-750.

[3] Cox D.R. y Oakes D. (1984). Analysis of Survival Data. Chapman & Hall, Great Britian, 1984

[4] Collett D. (1994). Modelling Survival Data in Medical Research. Chapman & Hall,Florida. 1994

[5] Chen H.Y. y Little R.J.A. (1999). Proportional hazards regression with missing covariates. Journal of the American Statistical Association.94, No.447: 896-908.

[6] Clark T.G. y Altman D.G. (2003). Developing a prognostic model in the presence of missing data: An ovarian cancer case study. J Clinical Epidemiology. 56: 28-37.

[7] Lawless J.F. (2003). Statistical models and methods for lifetime data. 2nd ed. John Wiley & Sons, Inc., New York.2003

[8] Rubin D. B. y Little R.J.A. (2002). Statistical analysis with missing data.2nd ed.John Wiley & Sons, Inc., New York. 2002

[9] Stephenson R.A. (2002). Prostate cancer trends in the era of prostate-specific antigen. An update of incidence, mortality, and clinical factors from the SEER database. The Urologic Clinics of North America. 29: 173-181.

[10] Van Buuren S., Boshuizen H.C. y Knook, D.L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine; 18: 681-694.

[11] Yang Y. Multiple Imputation for Missing Data: Concepts and New Development.Rockville, MD: SAS Institute, Inc. (Paper 267-25). (http://support.sas.com/rnd/app/papers/).

[12] Wayman J.C. (2003).Multiple Imputation For Missing Data: What Is It And How Can I Use It?.American Educational Research Association, Chicago, IL.

[13] Fichman M. y Cummings J.N. (2003).Multiple imputation for missing data: Making the most of what you know. Organizational Research Methods.Carnegie-Mellon University Pittsburgh PA. 6: 282-308.

[14] Schafer J. y Graham J. (2002). Missing Data: Our View of the State of Art. Psychological Methods.Vol. 7, No 2:147-177.

[15] Wilks S.S. (1932). Moments and Distributions of Estimates of Population Parameters from Fragmentary Samples. Annals of Mathematical Statistics. Vol. 3, No. 3, pp 163-195.

[16] Buck S.F. (1960). A method of estimation of missing values of multivariate data suitable for use with an electronic computer. Journal of the Royal Statistical Society. Series B (Methodological) Vol. 22, No. 2, pp. 302-306

[17] Rubin D. (1987). Multiple Imputation For Nonresponse In Surveys. JOHN WILEY & SONS.United States.

[18] Vergouwe Y., Royston P., Moons G.M. y Altman D. (2009) Development and validation of a prediction model with missing predictor data: a practical approach. Jurnal of Clinical Epidemiology. Volume 63, Issue 2, pp. 205-214

[19] Haesook T. (2007). Cumulative Incidence in Competing Risks Data and Competing Risks Regression Analysis. Clin Cancer Res January 2007;13:559-565.

[20] Coviello V. y Boggess M. (2004). Cumulative incidence estimation in the presebnce of competing risks. The Stata Journal 4, Number 2, pp. 103-112

[21] Janssen K., Donders R., Harrell F.E., Vergouwe Y., Chen Q., Grobbee D. y Moons K. (2010). Missing covariate data in medical research: To impute is better than to ignore. Journal of Clinical Epidemiology. Vol.6, pp. 721 - 727

[22] Vergouwea Y., Royston P., Moons K. y Altman D. (2010). Development and validation of a prediction model with missing predictor data: a practical approach. Journal of Clinical Epidemiology. Vol. 63, pp. 205 - 214.

[23] Van Buuren S. (2012). Flexible Imputation of Missing Data. Chapman and Hall/CRC, Boca Raton, FL.

[24] Farewell V.T. (1977). A model for a binary variable with time-censored observations. Biometrika. Vol. 64, No. 1, pp. 43-46

[25] Raghunathan T. E., Lepkowski J. M., Van Hoewyk, J. y Solenberger P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. Survey Methodology, Vol. 27, No 1.

[26] Aalen O.O. (1978). Nonparametric inference for a family of counting processes. Ann.Statist., Volume 6, Number 4, pp. 701-706

[27] Scrucca L., Santucci A. y Aversa F. (2010). Regression modeling of competing risk using R: an in depth guide for clinicians. Bone Marrow Transplantation. Vol. 45, 1388-1395

[28] Casella G. y Edward I.G. (1992).Explaining the Gibbs Sampler.The American Statistician Vol. 46, No. 3, pp. 167-174 Published by: American Statistical Association.URL: http://www.jstor.org/stable/2685208

[29] Dawid A. P. (1979). Conditional Independence in Statistical Theory. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 41, No. 1, pp. 1-31

[30] Gray R.J. (1988). A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk.The Annals of Statistics, Vol. 16, No. 3, pp. 1141-1154

[31] Gelman A.,King K. y Chuanhai L. (1999). Not Asked and Not Answered: Multiple Imputation for Multiple Surveys. Journal of the American Statistical Association. Vol. 93, pp. 846-857.

[32] Rubin D.B. (1987). Inference and missing data. Statistical analysis, Volume 63, Issue(3): 581-592.

[33] Chiu-Hsieh H., Taylor J. y Murray S. (2002). Survival estimation and testing via multiple imputation.Statistics & Probability Letters. Vol. 58, pp. 221-232

[34] Heymans M., Van Buuren S., Knol D.L., Van Mechelen W. y Henrica CW (2007). Variable selection under multiple imputation using the bootstrap in a prognostic study.BMC Medical Research Methodology. Vol. 7.

[35] Vergouwea Y., Royston P., Moons K. y Altmanc D. (2010). Development and validation of a prediction model with missing predictor data: a practical approach.Journal of Clinical Epidemiology. Vol. 63, pp. 205-214

[36] Royston P., White I.R. (2011). Multiple Imputation by Chained Equations (MICE): Implementation in Stata.Journal of Statistical Software. Volumen 45, Issue 4.

[37] White I.R., Royston P. y Wood A.M. (2010). Multiple imputation using chained equations: Issues and guidance for practice.Statistics Medicine. Vol. 30, Issue 4, pp. 377 - 399

[38] Trivellore R. y Bondarenko I. (2007). Diagnostics for Multiple Imputations . Available at SSRN: http://ssrn.com/abstract=1031750 or http://dx.doi.org/10.2139/ssrn.1031750

[39] Van Buuren S. y Oudshoorm K. (1999). Flexible multivariate imputation by MICE. Technical Report PG/VGZ/99.054, TNO Prevention and Health. Leiden, Holanda

[40] BRAND J.P.L. (1999). Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets. Thesis University of Rotterdam/TNO Prevention and Health.

[41] Kattan M.W., Heller G. y Brennan M.F. (2003). A competing-risks nomogram for sarcoma-specific death following local recurrence. Statistics in Medicine. Vol. 22, Issue 22, pp. 3515-3525

[42] Carpenter J.R. y Kenward M.G. (2013) Multiple Imputation and its Application. John Wiley & Sons Ltd. Chichester, UK.

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA

ACTA DE EXAMEN DE GRADO

No. 00097
Matrícula: 2113802147

IMPUTACION MULTIPLE PARA RIESGOS COMPETITIVOS: ANALISIS DE SUPERVIVENCIA PARA PACIENTES CON CANCER DE PROSTATA

En México, D.F., se presentaron a las 11:00 horas del día 14 del mes de noviembre del año 2013 en la Unidad Iztapalapa de la Universidad Autónoma Metropolitana, los suscritos miembros del jurado:

DR. GABRIEL ESCARELA PEREZ
DR. GABRIEL ARCANGEL RODRIGUEZ YAM
DRA. BLANCA ROSA PEREZ SALVADOR

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

MAESTRO EN CIENCIAS (MATEMÁTICAS APLICADAS E INDUSTRIALES)

DE: MARCO ANTONIO BARRAGAN MARTINEZ

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

APROBAR

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

MARCO ANTONIO BARRAGAN MARTINEZ
ALUMNO

REVISÓ

LIC. JULIO CESAR DE LARA ISASSI
DIRECTOR DE SISTEMAS ESCOLARES

DIRECTOR DE LA DIVISIÓN DE CBI

DR. JOSE ANTONIO DE LOS REYES HEREDIA

PRESIDENTE

DR. GABRIEL ESCARELA PEREZ

VOCAL

DR. GABRIEL ARCANGEL RODRIGUEZ YAM

SECRETARIA

DRA. BLANCA ROSA PEREZ SALVADOR