TRABAJO DE FIN DE MÁSTER

Máster Interuniversitario en Estadística e Investigación Operativa UPC-UB

Título: El modelo aditivo de Aalen. Una alternativa al modelo de riesgos proporcionales

Autor: Itxaso Alayo Bueno

Director: Klaus Langohr

Co-Director: Guadalupe Gómez

Departamento: Departamento de Estadística e Investigación Operativa

Universidad: Universitat Politècnica de Catalunya

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

UNIVERSITAT DE BARCELONA

El modelo aditivo de Aalen. Una alternativa al modelo de riesgos proporcionales

Máster en Estadística e Investigación Operativa

Universitat Politécnica de Catalunya

Autora: Itxaso Alayo Bueno

Director: Klaus Langohr
Co-directora: Guadalupe Gómez Melis

II

Resumen

Los modelos de regresión para datos de supervivencia han estado tradicionalmente basados en el modelo de Cox, el cuál asume que los riesgos son proporcionales. Como alternativa existe el modelo aditivo de Aalen que considera que los efectos de las covariables pueden variar a lo largo del tiempo. El objetivo de este trabajo es comparar estos dos modelos en pacientes diagnosticados de leucemia aguda en Estados Unidos entre el año 1973 y 2012. Los resultados muestran que cuando los riesgos proporcionales no se cumplen, el modelo aditivo de Aalen es una buena alternativa al modelo de Cox. En cambio, si la premisa de los riesgos proporcionales se cumple, ambos modelos pueden ser apropiados y, cada uno con la información que aporta, complementarios.

Palabras clave: Modelo de Cox, Modelo aditivo de Aalen, Leucemia Aguda.

Abstract

The regression models for survival data have traditionally been based on the Cox model, which assumes proportional hazards. An alternative is the additive Aalen model, which considers that effects from covariates can vary over time. The aim of this study is to compare the Cox and the Aalen models in acute leukemia patients from the United States between 1973 and 2012. Results show that when the hazards are not proportional, the Aalen additive model is a good alternative. When hazards are proportional both models can be appropiate because both offer different information.

Keywords: Cox model, Aalen's additive model, Acute leukemia.

Índice general

Capítulo 1

Introducción

1.1. Motivación

El cáncer es una de las principales causas de morbilidad y mortalidad en todo el mundo, siendo la segunda causa de muerte, detrás de las enfermedades cardiovasculares.

En 2016 se estima que se diagnosticarán 1.685.210 nuevos casos de cáncer en Estados Unidos y 595.690 personas morirán por la enfermedad. En 2012 hubo 14 millones de nuevos casos y 8,2 millones de personas murieron por causas relacionadas con el cáncer. En los próximos 20 años se prevé que el número de nuevos casos aumente en un 70 % [1].

Dentro de los cánceres la leucemia es el décimo cáncer más común y se estima que en 2016 habrá 60.140 nuevos casos y 24.400 fallecerán por la enfermedad. Esta enfermedad ocurre más comúnmente en adultos mayores de 55 años pero es a su vez el cáncer más común entre los niños menores de 15 años.

En este trabajo final de máster se estudiará la supervivencia en la leucemia aguda mediante 2 métodos. Primero mediante el método más habitual y conocido que es el modelo de regresión de Cox y por otro lado un modelo que no se estudia en el máster de estadística e investigación operativa, que es el modelo aditivo de Aalen.

Estos estudios se harán sobre pacientes diagnosticados de leucemia aguda entre los años 1973 y 2012 en Estados Unidos.

1.2. Leucemia

El cáncer se debe a una mutación de algunas de las millones de células de las que está formado el cuerpo. Cuando las células envejecen o sufren daños son reemplazadas por unas nuevas, pero el problema surge cuando este proceso se descontrola y las células no mueren (proceso que se conoce como mutación) y se crean nuevas que el cuerpo no necesita. Estas células sobrantes crean lo que conocemos como tumor.

La leucemia es un tipo de cáncer que afecta a las células de la sangre (células sanguíneas), y consiste en el aumento incontrolado de células anómalas de la sangre. Estas se infiltran en la médula ósea, impidiendo la producción de las otras células sanguíneas, e invaden la sangre y otros órganos.

Las células que se encargan de transportar el oxígeno al resto de células se llaman glóbulos rojos, las encargadas de proteger al organismo de infecciones son los glóbulos blancos y las células encargadas de evitar hemorragias en caso de corte o herida son las plaquetas.

La médula ósea es un tipo de tejido que se encuentra en el interior de los huesos y está formada por lo que llamamos células madres, que al madurar originan los tres principales tipos de células sanguínea que hemos mencionado anteriormente. La médula ósea mantiene el número de los tres tipos de estas células sanguíneas, sustituyendo las que mueren. Además siempre que sea necesario, es capaz de producir células de manera más rápida.

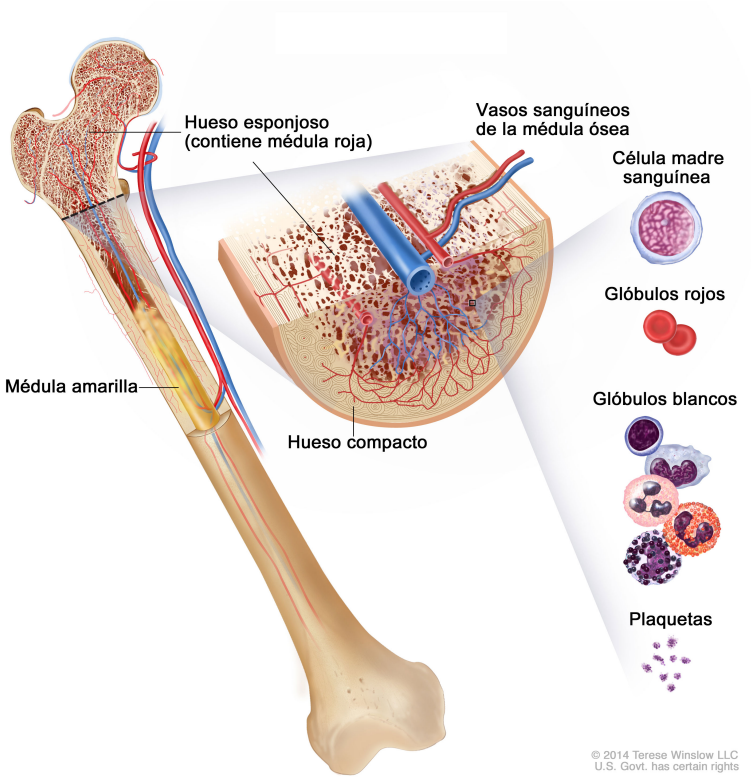

Figura 1.1: Anatomía del hueso [2].

1.2.1. Tipo de Leucemia

La leucemia tiene distintos tipos que se categorizan según la rapidez con la que avanza la enfermedad y según el tipo de célula que esté afectada.

Las leucemias agudas progresan muy rápidamente a consecuencia de un fallo severo en la función normal de la médula ósea. La salud de estos pacientes empeora rápidamente ya que los síntomas suelen ser anemia, infecciones o hemorragias. Este tipo de leucemias suelen requerir un ingreso hospitalario urgente en sala de aislamiento. Habitualmente es necesario realizar transfusiones para estabilizar los niveles sanguíneos.

En cambio las leucemias crónicas progresan lentamente y no suelen presentar síntomas, por lo que los pacientes suelen ser diagnosticados en análisis de sangre rutinarios. Al contrario que la leucemia aguda, en algunos casos se establecen controles estrictos en estos pacientes para vigilar la progresión de la enfermedad sin ser necesario comenzar un tratamiento. En las fases mas avanzadas, algunas leucemias crónicas pueden transformarse en leucemias agudas.

Los análisis de este trabajo estarán centrados en la leucemia aguda por lo que a continuación se explican los distintos tipos de esta que existen.

- La **leucemia linfocítica aguda (LLA)** o también llamada leucemia linfoblástica aguda se produce en un tipo de glóbulos blancos llamado *linfocitos*.

  Lo que ocurre en este tipo de leucemia es que hay demasiadas células que se convierten en linfoblastos, linfocitos B o linfocitos T (también llamadas células leucémicas) que no son capaces de combatir las infecciones de las que el enfermo puede infectarse. El aumento desproporcionado de estás células leucémicas hace que haya menos sitio para las células (glóbulos rojos, blancos y plaquetas) sanas.

  En ocasiones al diagnosticar por primera vez el cáncer, se encuentran linfocitos en la médula ósea y en ganglios linfáticos dificultando la diferenciación entre linfoma o leucemia. Si más del 25 % de la médula ósea es reemplazada por linfocitos cancerosos, usualmente la enfermedad se considera una leucemia.

- La **leucemia mieloide aguda (LMA)**, también llamada mielocítica aguda, habitualmente se origina porque células que se convertirían en glóbulos blancos (pero no en linfocitos) se transforman en glóbulos blancos inmaduros que se llaman mieloblastos. Esto hace que como ya hemos mencionado para el caso de la leucemia linfocítica aguda, haya menos espacio para los glóbulos blancos, glóbulos rojos y plaquetas sanas. La leucemia mieloide aguda se inicia en la médula ósea pero en la mayoría de los casos pasa rápidamente a la sangre.

- La **leucemia monocítica aguda** es una subcategoría de la condición más amplia de leucemia mieloide aguda (LMA). La leucemia monocítica se caracteriza específicamente por una sobreproducción de glóbulos blancos llamados monocitos y monoblastos, células que combaten las infecciones en el cuerpo.

1.2.2. Factores de riesgo y tratamiento

La exposición a radiación ionizante aumenta el riesgo de la mayoría de los tipos de leucemia. Una de las fuentes de radiación mas común es la radiación médica, como por ejemplo la que se utiliza en el tratamiento del cáncer. En muchas ocasiones la leucemia es consecuencia de la quimioterapia. También los niños con síndrome de Down u otro desorden genético tienen un riesgo mayor de padecerla. Los trabajadores de la industria de fabricación de caucho también tienen un riesgo mayor. Y hay estudios que sugieren que la obesidad aumenta el riesgo de leucemia.

Algunos factores de riesgo están más relacionados con un tipo específico de leucemia. Por ejemplo, los antecedentes familiares son un mayor factor de riesgo para algunas leucemias crónicas. El consumo del tabaco es un factor de riesgo para la LMA en adultos, y existen pruebas de que el tabaquismo de los padres/madres antes y después del parto, o la exposición durante el embarazo al tabaco puede aumentar el riesgo de leucemia infantil. Estar expuesto a ciertos productor químicos como el formaldehído o benceno (un componente en el humo del cigarrillo y la gasolina), aumenta el riesgo de padecer LMA.

No hay pruebas estándar para la detección precoz de la leucemia. Sin embargo, a veces se diagnostica pronto accidentalmente porque aparecen resultados anormales en las analíticas de la sangre que se realizan para otras enfermedades.

Respecto al tratamiento, para la mayoría de tipos de leucemia se utiliza quimioterapia. También se utilizan algunos medicamentos contra el cáncer, ya sea en combinación o como agentes únicos.

Algunos medicamentos dirigidos son eficaces para el tratamiento de leucemias crónicas, ya que atacan a las células con el Philadelphia chromosome , la anomalía genética característica de este tipo de leucemia. Pero algunos de estos medicamentos sirven también para el tratamiento de la leucemia linfocítica aguda, ya que este tipo puede implicar un defecto genético similar al de las leucemias crónicas.

En algunos casos y en condiciones apropiadas, para tratar cierto tipo de leucemias se utilizan altas dosis de quimioterapia seguidas de un trasplante de células madre. La médula del enfermo de leucemia es destruida mediante altas dosis de quimioterapia y reemplazada por una médula sana. Esta última puede proceder de un donante (trasplante alogénico), o bien del propio enfermo (trasplante autogénico o autólogo).

Teniendo en cuenta que algunas de las leucemias agudas actualmente no necesitan un trasplante alogénico y que los pacientes mayores de 65 años en principio no pueden someterse a un trasplante de estas características, es posible decir que solo un 20 % de enfermos con leucemias agudas requiere un trasplante alogénico.

1.2.3. Supervivencia y pronóstico

La supervivencia específica en cáncer es el porcentaje de los pacientes que no mueren por el cáncer durante un periodo de tiempo después del diagnóstico. Este periodo de tiempo puede ser de 1 año, 2 años, 5 años...etc, según el tipo de cáncer. En el caso de la leucemia el periodo

utilizado es de 5 años.

En el cáncer hay que diferenciar entre dos conceptos que son la curación y la remisión. Que un paciente se haya curado quiere decir que no quedan restos del cáncer después del tratamientos y el cáncer nunca volverá. Sin embargo el termino remisión significa que los signos y síntomas del cáncer han reducido. La remisión puede ser parcial o completa. En la completa todos los signos y síntomas del cáncer desaparecen. La mayoría de los cánceres que regresan lo hacen dentro de los primeros 5 años después del tratamiento.

Las tasas de supervivencia varían sustancialmente según el tipo de leucemia, que van desde una supervivencia relativa a los 5 años del 25 % de los pacientes diagnosticados con LMA al 84 % para las personas con LLA.

En el caso de la leucemia, para el cálculo de la supervivencia hay que tener en cuenta que hay pacientes que mueren por causas distintas a la enfermedad. Dentro de las causas de muerte hay algunas que son atribuibles a la leucemia, como por ejemplo las enfermedades infecciosas ya que una de las consecuencias de padecer leucemia es la disminución de las defensas.

Los avances en el tratamiento han dado lugar a una notable mejora en la supervivencia en las últimas tres décadas para la mayoría de los tipos de leucemia. Por ejemplo, del año 1975 al 2010, la tasa de supervivencia global relativa a los 5 años para la leucemia linfocítica aguda aumentó del 41 % al 70 %. Esta mejoría en la supervivencia se debe en gran parte al descubrimiento de fármacos dirigidos contra el cáncer, como por ejemplo el *imatinib*.

1.3. Objetivo del trabajo

El modelo de Cox, basado en la suposición de que las funciones de riesgo entre individuos expuestos a factores distintos se mantiene constante a lo largo del tiempo, es el modelo de regresión más utilizado para tiempos de supervivencia. Existe un modelo llamado modelo aditivo de Aalen y que es una alternativa al modelo de Cox. cuando no se cumplen los riesgos proporcionales.

El principal objetivo de este trabajo final de máster será a partir de una gran base de datos del Instituto Nacional de Cáncer de Estados Unidos de pacientes diagnosticados de leucemia analizar la supervivencia mediante el modelo de Cox y comparar estos resultados con los obtenidos mediante el modelo aditivo de Aalen.

En el primer capítulo se hace una breve introducción del cáncer, la leucemia, sus factores de riesgo y supervivencia. En el Capítulo 2 se explica el origen de la base de datos, las variables que contiene y los cambios realizados para hacer el estudio. Los métodos que se han utilizado, que son el modelo de Cox y el modelo aditivo de Aalen están descritos en el tercer Capítulo. En el Capítulo 4 se muestran los resultados de los modelos y la comparación entre ellos. Y por último en el Capítulo 5 se describe una breve discusión y conclusiones del trabajo.

Capítulo 2

Gestión de la base de datos

2.1. Historia

La base de datos utilizada proviene del Programa de Vigilancia, Epidemiología y Resultados (Surveillance, Epidemiology, and End Results Program; SEER)[3] del Instituto Nacional del Cáncer de Estados Unidos (National Cancer Institute; NCI) [2]. Este instituto trabaja para proporcionar información sobre las estadísticas de cáncer, en un esfuerzo para reducir la carga de cáncer entre la población de EEUU.

SEER recopila y publica la incidencia de cáncer y de los datos de supervivencia de los registros de cáncer de base poblacional que cubren aproximadamente el 28 % de la población estadounidense. El sitio web del programa SEER (`http://seer.cancer.gov/`) tiene estadísticas de cáncer más detalladas, incluidas las estadísticas de población para los tipos comunes de cáncer.

El Informe Anual sobre el Estado del Cáncer ofrece una actualización anual de la incidencia del cáncer, la mortalidad y las tendencias en Estados Unidos. Este informe se elabora conjuntamente por expertos del Instituto Nacional del Cáncer, los Centros para el Control y Prevención de Enfermedades, la Sociedad Americana del Cáncer y la Asociación de Registros Centrales del Cáncer de América del Norte.

2.2. Lectura de los datos

Para obtener la base de datos fue necesario el registro en SEER y firmar un acuerdo sobre el uso que se le daría a los datos. Una vez obtenidos los permisos se recibió un archivo en formato zip que contenía los registros de todos los pacientes diagnosticados de cáncer en Estados Unidos desde el año 1973 hasta el año 2012.

El archivo descargable contiene un archivo tipo texto para cada tipo de cáncer. Dichos archivos cubren los siguientes tipos de cáncer: pulmón, mama, colon y recto, otro tipo de digestivos, genitales femeninos, genitales masculinos, linfoma-mieloma-leucemia, respiratorio y tracto urinario.

Este trabajo se ha restringido a la leucemia y por este motivo se escoge únicamente el archivo

que contiene los datos de linfoma, leucemia y mieloma. Este archivo tipo texto contiene una fila por individuo pero las variables no están identificadas, lo que contiene es una fila con números y letras que son los valores de cada variable pero sin separar.

Una vez leída la base de datos lo primero de todo es separar los tres tipos de cáncer y seleccionar solo los pacientes de leucemia. Como ya se ha mencionado en el Capítulo 1, solo se analizará el tipo de leucemia aguda por lo que se seleccionan solo los pacientes diagnosticados con este tipo de leucemia.

A continuación se asignan los valores que corresponden a datos faltantes de cada variable según la información que se tiene de cada una de ellas. Por último se eliminan las filas donde no hay valores de supervivencia.

La base de datos inicial contenía 367088 individuos y 146 variables. Se eliminan 131 variables por contener más de el 50 % de valores perdidos o todos los valores en una única categoría. Nuestro estudio se restringe a la base de datos con 50572 individuos diagnosticados de leucemia aguda y 15 variables. Algunas de estas 15 variables únicamente dan información descriptiva, por lo que la dimensión de la base final que se utilizará para los análisis consta de 50572 individuos y 6 variables.

Para la lectura de bases de datos de SEER existe un paquete en R llamado SEER2R [16].
Junto con los datos había un documento pdf con la explicación de cada variable e indicación de longitud y posición de estas. Al intentar leer la base en R se ha visto que da error, debido a que las posiciones indicadas no son correctas ya que hay algunas columnas vacías de más. Gracias a un editor de texto se eliminan las columnas sobrantes y la base finalmente se ha leído con el comando de R *read.fwf*.

A continuación, la Figura 2.1 muestra el diagrama de flujo con todos los pasos explicados anteriormente para la lectura de los datos.

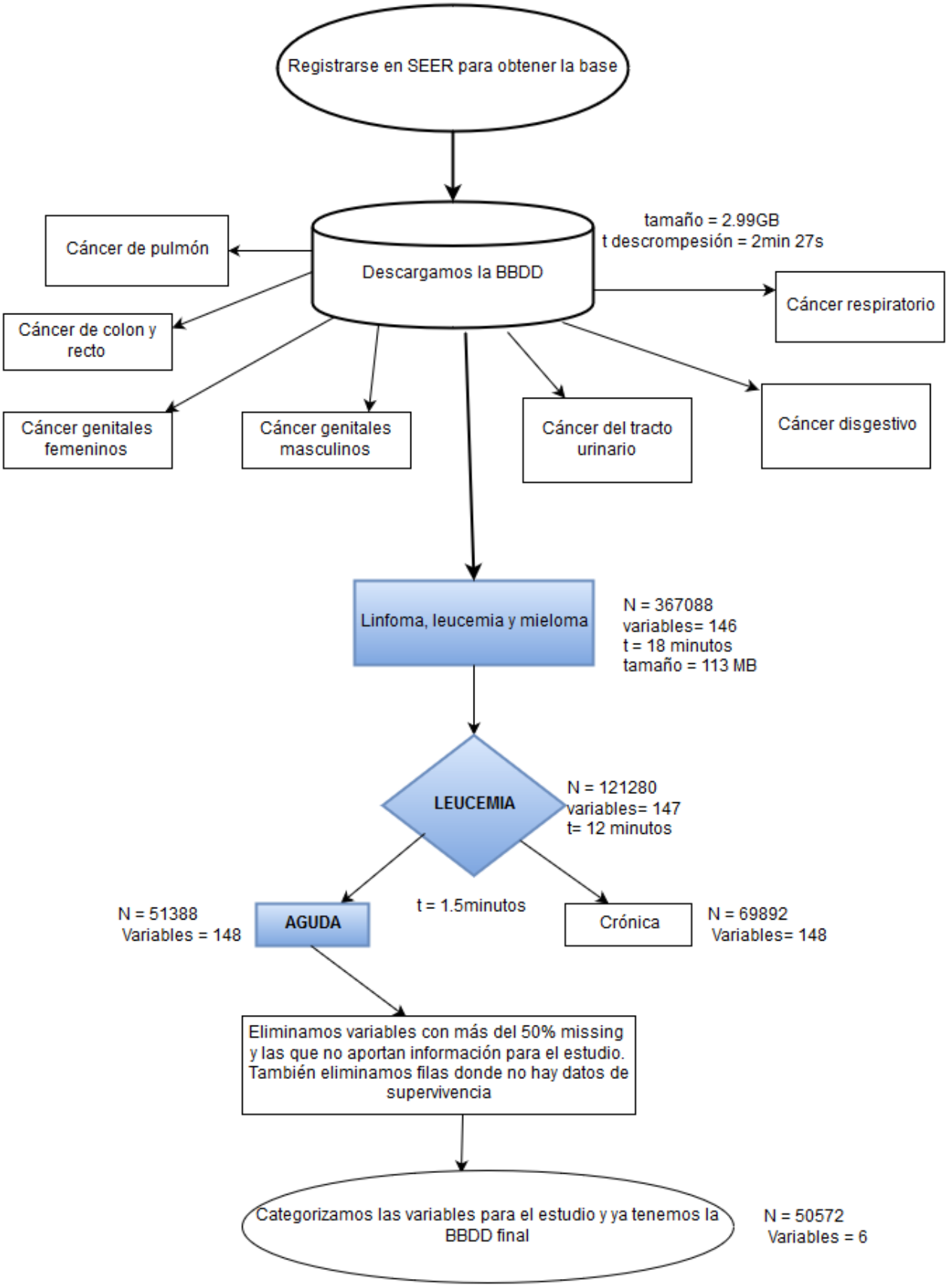

Figura 2.1: Diagrama de flujo

2.3. Variables

Se han quitado de la base 69 variables que contenían 100 % de missing, 19 variables que contenían más de un 60 % de missing y 3 variables que tenían un 50 % de missing. A continuación en la Tabla 2.1 se muestran las variables restantes.

Variable	Descripción	Missing (%)
adiag	Año del diagnóstico	0 (0 %)
anaci	Año de nacimiento	0 (0 %)
AYA	Se utiliza principalmente para analizar los datos sobre adolescentes y adultos jóvenes	0 (0 %)
causa	Causa de la muerte (cancerosa o no)	0 (0 %)
cemuerte	Designa si el paciente murió por su cáncer	7520 (14.64 %)
comp_o2	Código comportamiento ICD-O-2	0 (0 %)
comp_o3	Comportamiento (92-00) ICD-O-3	0 (0 %)
comp_recod	Creado para poder eliminar datos que no fueron recogidos coherentemente con el tiempo	0 (0 %)
condado	Estado y condado del paciente	0 (0 %)
confir	Confirmación del diagnóstico	1040 (2 %)
Csesquema	La información de la CS se recoge bajo las especificaciones de un esquema concreto basado en localización e histología	0 (0 %)
DX	Estado civil en el momento del diagnóstico	1625 (3.17 %)
ediag	Edad en el momento del diagnóstico	3 (0.05 %)
ediag_recod	Nuevo código de edad, basado en la edad en el diagnóstico	0 (0 %)
EOD	Extensión de la enfermedad	14163 (27.61 %)
EOD_13	Para seleccionar los sitios de los tumores diagnosticados	0 (0 %)
follow_up	Este código cifra el tipo de seguimiento previsto para un caso SEER	0 (0 %)
fuente	Fuente del caso	0 (0 %)
hisp	Tipo de hispano	0 (0 %)
hisp_recod	Recodificación de la variable hisp	0 (0 %)
hist_broad	Histología (usada para separar leucemia, linfoma y mieloma)	0 (0 %)
hist_o2	Tipo histológico del tumor según ICD-O-2	0 (0 %)
hist_o3	Tipo histológico del tumor según ICD-O-3	0 (0 %)

Variable	Descripción	Missing (%)
ICD_10	Sitio primario y morfología según ICD-10	0 (0 %)
ICD_3_ext	Recodificación de la histología según ICD-O-3	0 (0 %)
ICD_9	Sitio primario y morfología según ICD-9	0 (0 %)
id	Población donde se hizo el registro	0 (0 %)
id_num	Número identificador del paciente	0 (0 %)
IHS	Nativos americanos	10581 (21 %)
KM	Recodificación basada en una causa básica de defunción para designar la causa de muerte en grupos similares	0 (0 %)
lat	Lado del órgano o del cuerpo donde se origina el tumor	0 (0 %)
maligno	Si es o no maligno el tumor	0 (0 %)
mdiag	Mes en el que el tumor fue diagnosticado por primera vez	0 (0 %)
msurv	Supervivencia (en meses)	816 (1.5 %)
msurv_flag	Recodificación de la supervivencia	0 (0 %)
msurv_fvivo	Recodificación de la supervivencia	0 (0 %)
msurv_vivo	Estado vital	816 (1.5 %)
num	Número de todos los tumores primarios malignos,in situ, benignos, y límite, que se producen durante la vida de un paciente.	0 (0 %)
num_prim	Número de todos los tumores primarios (basado en el número total de tumores en SEER)	0 (0 %)
opera	Razón de porqué la cirugía no se realizó en el sitio primario	3176 (6 %)
origen	Pacientes con apellido español / hispano o de origen español	0 (0 %)
other	Designa que el paciente murió por causas ajenas a su cáncer	7520 (14.64 %)
prim	Sitio donde se originó el tumor primario.	0 (0 %)
prim_inter	Primario por normas internacionales	0 (0 %)
prim_recod	Recodificación basada en sitio primario e histología ICD-O-3 del tumor	0 (0 %)
rad_oper	Orden en el que se administraron la cirugía y radiación para los pacientes que tenían las dos terapias.	0 (0 %)
radio	Método de la terapia de radiación realizado como parte del primer curso de tratamiento.	849 (1.6 %)
raza	Raza	0 (0 %)

Variable	Descripción	Missing
raza_recod	Nuevo código de raza (basado en las variables de raza e IHS)	154 (0.3 %)
raza_recod2	Recodificación de la variable raza	154 (0.3 %)
SEER	Número de registros que se han presentado a SEER para cada paciente	0 (0 %)
SEER_A	Es una versión simplificada de etapa: in situ, localizado, regional, distante, y desconocido	0 (0 %)
sex	Género	0 (0 %)
vital	Estado vital del paciente	0 (0 %)

Tabla 2.1: Variables con menos del 50 % de valores perdidos. En el anexo se muestras los códigos oncológicos mencionados en la Tabla.

2.4. Adaptación de la base de datos

Para llevar a cabo el estudio ha sido necesario la modificación de variables ya existentes o la creación de variables nuevas. Partiendo de la variable *hist_broad* se crea una variable nueva llamada *grupo*, con la que se diferencia entre las tres enfermedades que contiene la base de datos, linfoma, leucemia y mieloma.

A continuación desde la variable *prim_recod* se crean dos variables, por un lado la variable *tipo* que nos servirá para diferenciar entre leucemia aguda y crónica, y por otra lado la variable *tipo_aguda* para diferenciar distintos tipos dentro de la leucemia aguda.

La variable continua *ediag* ha sido categorizada en cuatro categorías. Estas categorías se han hecho según criterio medico ya que los tratamiento de la leucemia se aplican según la franja de edad en la que está el paciente en el momento del diagnóstico.

Otra de las variables continuas categorizada ha sido *adiag*. Estas categorías se han hecho por décadas, que es en los años en los que se han visto cambios en los tratamientos.

Respecto a la variable *msurv* he de decirse que se ha supuesto que los pacientes morían en mitad del mes ya que esta variable recoge la supervivencia en meses, sin especificar días, y muchos de los pacientes de leucemia no sobreviven al primer mes lo que causaba problemas en algunos análisis por concentrase un gran número de fallecidos en el momento cero.

Por último tenemos la variable *causa* en la que están codificadas todas las causas de muerte de los pacientes. Esta variable se ha utlizado para crear la *censura*. Primero de todo se ha establecido el tiempo de supervivencia en 5 años, es decir, todos los pacientes que no han fallecido en los primeros 5 años desde el diagnóstico se considerará que están curados. Después se ha considerado como valor 1 los individuos muertos por leucemia aguda u otras enfermedades que están relacionadas con esta enfermedad, es decir, causa por las que los pacientes han fallecido

por estar enfermos de leucemia y que en otro caso no hubieran fallecido. Estas enfermedades son las siguientes: tuberculosis, sífilis, septicemia, otras infecciones y enfermedades parasitarias, neumonía y gripe, y ulceras estomacales y duodenales. Y con 0 se han categorizado todos los individuos que están vivos dentro de los 5 años después del diagnóstico, los que han superado los 5 años desde el diagnóstico y los pacientes que han muerto por causas ajenas a la enfermedad.

En la Tabla 2.2 aparece una descripción de las variables utilizadas en el estudio.

Variable	**Categoría**	**Frecuencia (%)**
Histología	Malignant lymphomas	31526 (8,6 %)
	Hodgkin lymphomas	28913 (7,8 %)
	Nhl - mature b-cell lymphomas	119281 (32,5 %)
	Nhl - mature t and nk-cell lymphomas	11395 (3,1 %)
	Nhl - precursor cell lymphoblastic lymphoma	1399 (0,38 %)
	Plasma cell tumors	52447 (14,3 %)
	Mast cell tumors	13 (0,003 %)
	Leukemias	7164 (1,95 %)
	Lymphoid leukemias	57811 (15,75 %)
	Myeloid leukemias	50433 (13,74 %)
	Other leukemias	5702 (1,55 %)
	Chronic myeloproliferative disorders	170 (0,04 %)
Enfermedad	Linfoma	192527 (52,44 %)
	Leucemia	121280 (33,04 %)
	Mieloma	52447 (14,3 %)
	Otras	834 (0,22 %)
Histología de leucemia	Leucemia linfocítica aguda	12552 (10,35 %)
	Leucemia linfocítica crónica	42685 (35,2 %)
	Otra leucemia linfocíticas	4460 (3,68 %)
	Leucemia mieloide aguda	32305 (26,64 %)
	Leucemia mieloide crónica	16215 (13,37 %)
	Otra leucemia monocítica/mieloide	2383 (1,69 %)
	Leucemia monocítica aguda	2216 (1,83 %)
	Otras leucemias agudas	4315 (3,56 %)
	Otras leucemias	3690 (3,04 %)
Tipo de leucemia	Leucemia Aguda	51388 (42,37 %)
	Leucemia Crónica	58900 (48,57 %)
	Otras	10533 (8,68 %)
Tipo de leucemia aguda	Linfocítica	12492 (24,7 %)
	Mieloide	31863 (63 %)
	Monocítia	2198 (4,35 %)
	Otras	4019 (7,95 %)

Variable	Categoría	Frecuencia (%)
Edad al diagnóstico	Media	52,13
	Mediana	61
	Mínimo	0
	Máximo	111
Edad al diagnóstico categorizada	< 14	8289 (16,39 %)
	[14,45)	8856 (17,51 %)
	[45,65)	10941 (21,64 %)
	$\geq$65	22483 (44,46 %)
Año del diagnóstico	Media	1995
	Mediana	1996
	Mínimo	1973
	Máximo	2012
Año del diagnóstico categorizada	[1973,1980)	6148 (12,16 %)
	[1980-1990)	11053 (21,86 %)
	[1990-2000)	13279 (26,26 %)
	[2000-2012]	20092 (39,73 %)
Género	Hombre	27661 (54.7 %)
	Mujer	22911 (45.3 %)
Supervivencia en meses	Media	36 meses
	Mediana	7 meses
	Mínimo	0
	Máximo	60 meses
Causa de la muerte	Vivos	13270 (26,24 %)
	Muerte por causa no atribuible a la leucemia aguda	8277 (16,37 %)
	Muerte por leucemia aguda o causas atribuibles	29025 (57,39 %)
Censura	Censurado	21547 (42,61 %)
	No censurado	29025 (57,39 %)

Tabla 2.2: Variables utilizadas en el estudio.

Capítulo 3

Análisis de supervivencia

En supervivencia los modelos de regresión normalmente han estado basados en el modelo de regresión de Cox. Sin embargo la validez de este modelo depende de la aceptación de los riesgos proporcionales. Además otra de las restricciones que tiene este modelo es que asume que los coeficientes de regresión son constantes en el tiempo. Una alternativa al modelo de Cox es el modelo aditivo de Aalen, que permite analizar los coeficientes como variables en el tiempo.

3.1. Tiempo de supervivencia

Los métodos de supervivencia son muy usados en investigación en áreas de la salud. La mayor diferencia entre los datos de supervivencia y otros datos numéricos continuos, es que es posible que no se observe el valor exacto de la variable de interés.

En este trabajo el evento de interés será la muerte por leucemia aguda y el tiempo T, que se conoce como tiempo de supervivencia, será el tiempo que transcurre desde que el paciente es diagnosticado hasta que muere por la enfermedad dentro de los primeros 5 años desde el diagnóstico (ya que como se explica en el Capítulo 1, a partir de los 5 años desde el diagnóstico se considera que el paciente está curado).
La variable aleatoria T es no negativa y corresponde a una población homogénea.

La función de densidad de T es el límite de la probabilidad de un paciente de morir por la enfermedad en un intervalo de tiempo $[t, t+\Delta]$ por unidad de tiempo y se representa por

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < \Delta t)}{\Delta t}$$

La función de distribución de T es la probabilidad de que la muerte ocurra antes de t y se representa por

$$F(t) = \Pr(T \leq t)$$

A la probabilidad de que un paciente sobreviva más de t unidades de tiempo se le llama **función de supervivencia** y se denota de la siguiente forma:

$$S(t) = \Pr(T > t) = 1 - F(t) = \int_t^{\infty} f(s)ds \quad , \; t \geq 0$$

3.1.1. Censura

A veces ocurre que no se puede observar el tiempo en el que ha ocurrido el evento de interés, que en este caso es la muerte por leucemia aguda, para algún individuo. Hay distintas razones por las que esto puede ocurrir, puede ser porque el paciente ha salido del estudio antes de los 5 años desde la fecha del diagnóstico, porque ha muerto por una causa ajena a la leucemia aguda, como puede ser por ejemplo un accidente de tráfico, o bien porque en el año 2012 que fue el último año de recogida de datos aún no había muerto, pero no habían pasado 5 años desde el diagnóstico, por lo que no se sabe que hubiese pasado con ese paciente.
A este tipo de casos se les llama datos **censurados**, concretamente censurados por la derecha.

En general, para una muestra de n individuos con tiempos potenciales hasta la muerte denotados por $T_1, ..., T_n$, si tenemos censura observaremos los pares (Y_1,δ_1),...,(Y_n,δ_n) donde $Y_i = \text{mín}\{T_i, C_i\}$, siendo $C_1, ..., C_n$ los tiempos de censura para cada uno de los individuos y

$$\delta_i = \begin{cases} 1 \text{ si } T_i \leq C_i \text{ ; el individuo } i \text{ no está censurado} \\ 0 \text{ si } T_i > C_i \text{ ; el individuo } i \text{ está censurado} \end{cases}$$

Existen otros tipos de censura que no se presentan en este trabajo como son la censura por la izquierda, censura en un intervalo y la censura doble.

3.1.2. Función de Riesgo

En muchas situaciones es importante saber como el riesgo de sufrir un concreto evento cambia en el tiempo, para ello existe la función de riesgo. Por ejemplo, se sabe que en el caso de la mortalidad infantil el riesgo más alto es en los días próximos al parto y luego decrece rápidamente.

La **función de riesgo** de T se puede interpretar como el comportamiento de la probabilidad condicionada de morir por leucemia aguda en un pequeño intervalo de tiempo sabiendo que el paciente estaba vivo al inicio de ese intervalo de tiempo.

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \Pr(t \leq T \leq t + \Delta t | T \geq t)$$

Cuando el riesgo es alto la supervivencia decae rápidamente, mientras que si el riesgo es cero la curva de supervivencia será plana.

En el caso continuo esto se puede expresar de la siguiente manera:

$$S(t) = \frac{f(t)}{\lambda(t)}$$

Otra cantidad relacionada es el **riesgo acumulado**. Aunque es muy útil gráficamente y también técnicamente, carece de interpretación intuitiva.

$$\Lambda(t) = \int_0^t \lambda(s) ds$$

Y cuando los datos son continuos se cumple:

$$S(t) = \exp\left(-\Lambda(t)\right) = \exp\left(-\int_0^t \lambda(s)ds\right)$$

Si la distribución del tiempo de supervivencia se asume que es exponencial, entonces la función de riesgo que hemos descrito anteriormente se puede estimar mediante la razón entre el número de eventos observados dividido por el total de tiempo de supervivencia. Una de las consecuencias de asumir la distribución exponencial es que el riesgo, λ, no varía en el tiempo. Un ejemplo de esto en nuestro estudio implica que el riesgo de morir por leucemia aguda es igual en los primeros meses después del diagnóstico que por ejemplo a los 4 años.

3.2. Modelo de regresión de Cox

El modelo de Cox (introducido por Cox en 1972), también llamado modelo de riesgos proporcionales, es el modelo más utilizado en supervivencia para cuando hay datos censurados. Este modelo es un equivalente a una regresión lineal pero para el ámbito de la supervivencia, y tiene parecido con una regresión logística de las tasas de riesgo.

Este modelo asume que la función de riesgo es constante sobre un periodo de tiempo y el efecto de las covariables se relaciona linealmente con el logaritmo de la razón de riesgos.

La función de riesgo del modelo de Cox tiene la siguiente expresión:

$$\lambda(t|\mathbf{X}) = \lambda_0(t)\ \exp\{\beta'\ \mathbf{X}\} = \lambda_0(t)\ \exp\{\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p\} \tag{3.1}$$

donde λ_0 es la función de riesgo basal (valor de la función de riesgo para aquellos individuos con valor 0 en las covariables, en nuestro caso será hombre, menor de 14 años y con leucemia linfocítica aguda diagnosticada entre los años 1973 y 1980). $X_1, ..., X_p$ son covariables fijas, es decir, las covariables explicativas género, tipo de leucemia aguda, edad al diagnóstico y año del diagnóstico. Al conjunto de todos los factores pronóstico o covariables asociadas a un individuo lo denotaremos por $\mathbf{X} = (X_1, X_2, ..., X_p)'$.

La estimación de los coeficientes β se hace mediante la maximización de la función de verosimilitud parcial. Esta función se denota por $L(\boldsymbol{\beta})$ y se define por

$$L(\beta_1, ..., \beta_p) = \prod_{j=1}^{r} \Pr\left(e_j = i|\Gamma_j\right) = \prod_{j=1}^{r} \Pr\left(X_{(j)} = x_{(j)}|\Gamma_j\right)$$

donde x_j es el vector de covariables, e_i indica a que individuo corresponde la muerte y $\Gamma = (Y_i, \delta_i, \boldsymbol{X}_i)$ es el conjunto que contiene toda la información de la muestra, y Γ_i es el conjunto de toda la información disponible hasta el momento $t_{(j)}$ sabiendo que ha ocurrido una muerte en $t_{(j)}$.

La función de verosimilitud parcial se puede expresar también como:

$$L(\beta_1, ..., \beta_p) = \prod_{j=1}^{r} \frac{\exp\{\beta' x_{(j)}\}}{\sum_{l \epsilon R(t_{(j)})} \exp\{\beta' x_{lj}\}} \tag{3.2}$$

o equivalente a

$$L(\beta_1, ..., \beta_p) = \prod_{i=1}^{r} \left(\frac{\exp\{\beta' x_{(i)}\}}{\sum_{l \epsilon R(Y_{(i)})} \exp\{\beta' x_{(i)}\}} \right)^{\delta_i}$$

Volviendo al modelo de Cox y a la función (3.1), se puede reescribir como:

$$\frac{\lambda(t|\mathbf{X})}{\lambda_0(t)} = \exp\{\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p\}$$

donde el término de la derecha sólo depende de los valores de las covariables y no del tiempo t.

El factor $\exp\{\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p\}$ corresponde al riesgo relativo de un paciente con perfil **X** respecto de un paciente con perfil **X=0**, y que como ya hemos mencionado anteriormente tenemos fijado en un paciente hombre con leucemia linfocítica aguda, menor de 14 años y diagnosticado entre el año 1973 y 1980.
Este factor expresa cuántas veces mayor es el riesgo instantáneo de padecer el evento de interés con un perfil **X** que con un perfil **X=0**.

3.2.1. Validez del modelo de Cox

La validez del modelo de Cox depende en gran medida de la validación de la proporcionalidad. Esta hipótesis se puede comprobar de forma analítica o de forma gráfica mediante los residuos.

Los residuos se calculan para cada paciente y proporcionan información sobre la diferencia entre el valor de supervivencia observado y el valor estimado por la ecuación de regresión. Cuanto mayor es esa diferencia mayor será el valor del residuo.

Residuos

Existen distintas definiciones de residuos, están los basados en martingalas, los residuos basados en la *deviance*, los residuos basados en el *score* y los residuos de *Schoenfeld*.

A continuación se explican lo residuos de *Schoenfeld*, que son los que se utilizan en este trabajo para validar el modelo, ya que se distribuyen aleatoriamente para cada covariante que cumpla el modelo de Cox.

La expresión de los residuos de *Schoenfeld* para el i-ésimo individuo y la k-ésima covariable, es la siguiente:

$$r_{SC_{ik}}(t) = \delta_i J_i(t)\{X_{ik} - \bar{X}_k(T_i)\}$$

donde $J_i(t)$ es el indicador de que el individuo j está en el estudio en el momento justo antes del tiempo t, δ_i es el indicador de censura, X_{ik} es el valor de la k-ésima covariable del individuo i y $\bar{X}_k(T_i)$ es el valor promedio de la covariable k en el tiempo T_i .

Cada uno de estos residuos establece la diferencia entre el valor observado de la covariante en el k-ésimo tiempo de fallo y el valor esperado de la covariante en ese momento.

Comprobación de la proporcionalidad gráficamente

Se podrá decir que el modelo cumple la premisa de proporcionalidad si la estimación de los parámetros se mantiene constante a lo largo del tiempo.

Para ver gráficamente que se cumple esta premisa para una covariable, los residuos deben agruparse de forma aleatoria a ambos lados del valor 0 del eje Y, y la curva del estimador β de esa covariable debe ser una línea recta con pendiente cero (línea discontinua), como se puede ver en el ejemplo de la Figura 3.1.

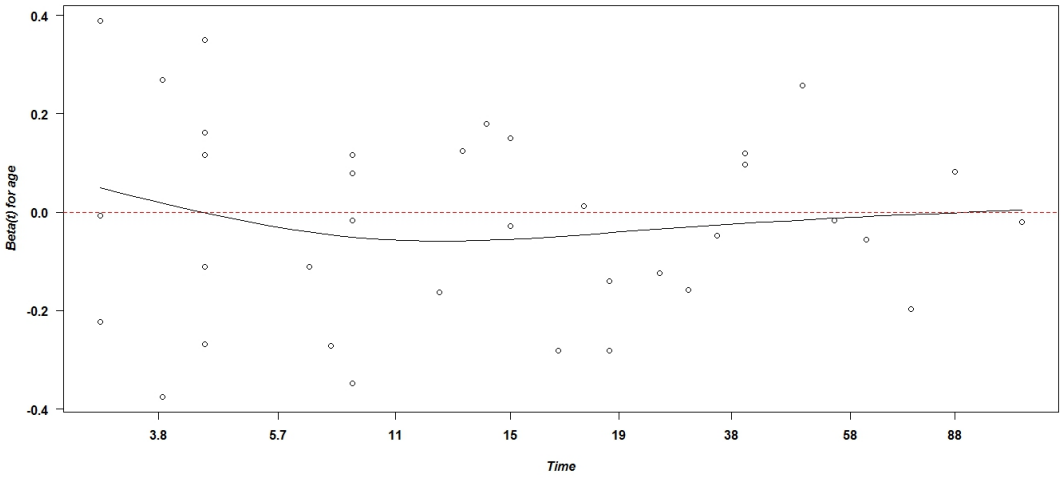

Figura 3.1: Ejemplo de residuos de Schoenfeld de una covariable que cumple los riesgos proporcionales. La curva del estimador de la covariable (línea continua) se ajusta a la línea recta de pendiente cero (línea discontinua).

3.2.2. Modelo de Cox estratificado

Como se ha visto en el apartado anterior puede ocurrir que la hipótesis de riesgos proporcionales no se cumpla para alguna covariable. En este caso una posibilidad es estratificar por esa covariable y hacer el modelo de riesgos proporcionales dentro de cada estrato para las demás covariables.

En estos casos los individuos del estrato j-ésimo tienen función de riesgo subyacente distinta, aunque el efecto de las otras covariantes en la supervivencia es el mismo.

El conjunto de datos observados se resume mediante

$$D = \{Y_i, \delta_i, e_i, [X_i(t), 0 \leq t \leq Y_i], i = 1, 2, ..., n\}$$

donde Y_i es el tiempo en estudio del paciente i-ésimo, δ_i es el indicador de censura, X_i es el vector de covariables medido mientras el individuo ha estado en el estudio y e_i es el estrato al que pertenece el individuo. Supondremos que e_i es una variable categórica que toma s valores.

El individuo del j-ésimo estrato tiene una función de riesgo basal arbitraria, $\lambda_{0j}(t)$, y el efecto de las otras covariables explicativas en la función de riesgo pueden ser representadas por el modelo de riesgos proporcionales en este estrato como:

$$\lambda_j(t|\mathbf{X}(t)) = \exp\{\beta'\mathbf{X}(t)\}\ \lambda_{0j}(t) = \exp\{\beta_1 X_1(t) + \beta_2 X_2(t) + ... + \beta_p X_p(t)\}\ \lambda_{0j}(t)$$

En este modelo se asume que el coeficiente de regresión es el mismo en cada estrato aunque las funciones de riesgo basal pueden ser diferentes y no tienen porque estar relacionadas entre ellas.

La estimación y el test de hipótesis se hace igual que en el modelo de Cox sin estratificar, pero ahora la función de verosimilitud parcial es la siguiente:

$$L(\beta) = L_1(\beta)\ \text{x}\ L_2(\beta)\ \text{x}\ \cdots\ \text{x}\ L_s(\beta)$$

donde $L_j(\beta)$ es la verosimilitud parcial para el estrato j, como en la fórmula (3.2) pero usando solamente los individuos del estrato j-ésimo.

3.3. Modelo aditivo de Aalen

En el modelo de riesgos proporcionales explicado en el apartado anterior, el efecto de las covariables actúa de forma multiplicativa en la función de riesgo basal. Los coeficientes β son constantes en el tiempo y se estiman mediante la maximización de la función de verosimilitud parcial.

En este apartado se presenta un modelo alternativo en el que las covariables pueden ser funciones dependientes del tiempo. Además en este modelo las covariables actúan de manera aditiva en la función de riesgo basal (lo que supone que las covariables tienen efectos separados en la variable respuesta). Por este motivo se llama modelo de riesgos aditivos o modelo aditivo de Aalen, y fue introducido por Aalen en 1980.

3.3.1. Formulación del modelo

Como en este modelo las covariables pueden variar a lo largo del tiempo, los coeficientes pueden ser funciones dependientes del tiempo, por lo tanto la expresión del modelo no-paramétrico es la siguiente:

$$\lambda(t|\boldsymbol{X}(t)) = I(t)\boldsymbol{X}(t)^T\boldsymbol{\beta}(t) = I(t)(X_1(t)\beta_1(t) + ... + X_p(t)\beta_p(t)) \tag{3.3}$$

y su función de riesgo acumulado:

$$\Lambda(t|\boldsymbol{X}(s)) = \int_0^t \lambda(s|\boldsymbol{X}(s))ds = \int_0^t I(s)\boldsymbol{X}(s)^T\boldsymbol{\beta}(s)ds$$

donde $\boldsymbol{\beta}(t)$ es un vector de funciones integrables[1] de dimensión p, $\boldsymbol{X}(t)$ un vector de covariables de dimensión p e $I(t)$ es un indicador de riesgo que toma valor 0 o 1. Este indicador será 1 si el

[1]Sea f una función acotada definida en un intervalo [a,b] (cerrado y acotado). f es integrable si y sólo si para cada $\varepsilon > 0$ existe una partición P_ε de [a,b] tal que $S_{Sup}(f, P_\varepsilon) - S_{inf}(f, P_\varepsilon) < \varepsilon$.

paciente está a riesgo en el momento t y 0 en caso contrario.

Para entender mejor el significado y definición de que un coeficiente dependa de t, a continuación en la Figura 3.2 se muestra un ejemplo gráfico de los coeficientes de regresión acumulados de dos covariables. Los gráficos son de los los coeficientes de regresión acumulados ya que como se explica en el siguiente apartado, es mediante lo que se estimarán los coeficientes de las covariables

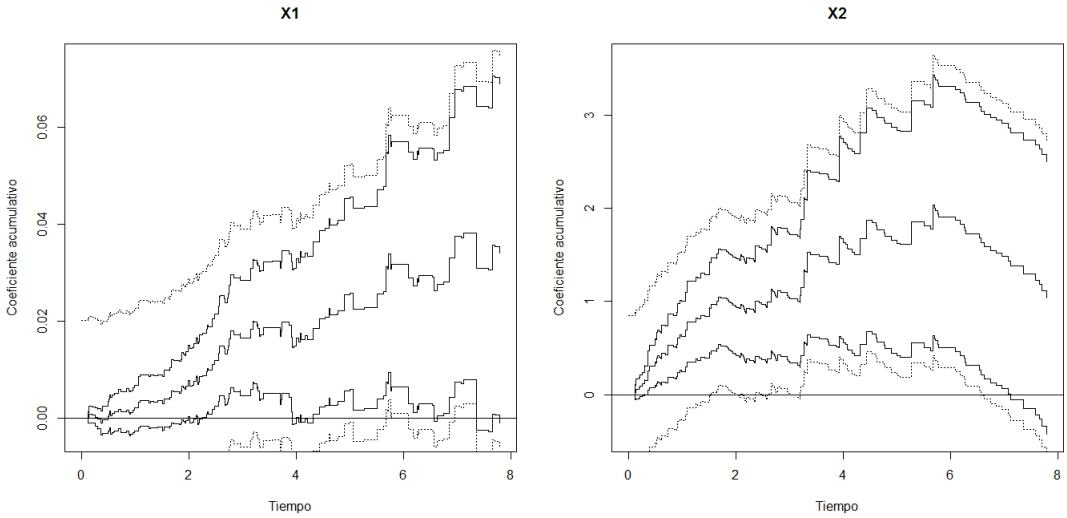

Figura 3.2: Estimación de la funcion de regresión acumulada del modelo aditivo de Aalen con intervalos de confianza del 95 % (líneas continuas) y bandas uniformes de confianza del 95 % (líneas discontinuas) para las variables X_1 y X_2.

En el gráfico de la estimación del coeficiente β de la covariable X_1 se puede ver una curva que crece constante a lo largo del tiempo (con algunas variaciones que pueden ser causa del volumen de los datos), lo que indica que el coeficiente β no varía en el tiempo.

Sin embargo en el gráfico de la covariable X_2 se observa que la curva no es una línea recta, es creciente hasta el tiempo 6 en el que comienza a decrecer. Por este motivo el coeficiente de X_2 no es constante, si no que es una función dependiente del tiempo.

3.3.2. Estimación de los parámetros

La estimación en el modelo aditivo de Aalen se centra en la función de regresión acumulada, ya que como los coeficientes β pueden ser funciones de t no es posible estimarlos mediante la función de verosimilitud.

En este caso los coeficientes se estiman calculando el estimador de la integral del coeficiente de regresión acumulada. La expresión del coeficiente de regresión acumulado es el siguiente:

$$B(t) = \int_0^t \beta(s)ds$$

Para entender de donde proviene el estimador de $B(t)$ que se va a calcular, se empieza dando unas breves nociones sobre los procesos contadores y el rol que estos juegan en la estimación de B(t).

Sea $N(t) = (N_1(t), ..., N_n(t))^T$ un proceso contador, $N(t) = I(T \leq t)$, y $\lambda(t) = (\lambda_1(t), ..., \lambda_n(t))^T$ su función de riesgo. Además la función de riesgo acumulada se denota como:

$$\Lambda(t) = \int_0^t \lambda(s)ds$$

Siendo $M(t) = N(t) - \Lambda(t)$ una martingala [2] de media cero, se puede escribir la derivada del proceso contador como:

$$dN(t) = dM(t) + d\Lambda(t) = dM(t) + \lambda(t)dt \tag{3.4}$$

A continuación se organizan las covariables en una matriz de diseño de dimensión n x p:

$$\mathbf{X}(t) = (Y_1(t)X_1(t), ..., Y_n(t)X_n(t))^T$$

Si se sustituye la ecuación (3.3) en la ecuación (3.4), se obtiene:

$$dN(t) = dM(t) + \lambda(t)dt = dM(t) + \boldsymbol{X}(t)\boldsymbol{\beta}(t)dt \tag{3.5}$$

y como los incrementos de la martingala no están correlacionados y su media es cero, la ecuación (3.5) sugiere que los incrementos de $\beta(t)$, que se describen como $dB(t)$, se pueden estimar mediante técnicas simples de regresión lineal múltiple (método de mínimos cuadrados).

Se define primero la inversa generalizada de $\boldsymbol{X}(t)$ como una matriz p x n

$$X^-(t) = (X^T(t)W(t)X(t))^{-1}X^T(t)W(t)$$

donde $W(t)$ es una matriz de pesos diagonal de dimensión n x n.

De la descomposición de la martingala anterior (ecuación (3.5)) se llega al estimador de $B(t)$ con el que se estimarán los coeficientes del modelo :

$$d\hat{B}(t) = X^-(t)dN(t)$$

que se puede escribir de la forma

$$\hat{B}(t) = \int_0^t X^-(s)dN(s).$$

Hay que tener en cuenta que

[2] Se dice que $M_0, M_1, ...$ es una martingala si para cualquier $n \geq 0$:

- $E|M_n| < \infty$
- Para cualquier sucesión de posibles valores $m_0, m_1, ..., m_n$
  $E[M_{n+1}|M_0 = m_0, M_1 = m_1, ..., M_n = m_n] = m_n$

$$\hat{B}(t) = \int_0^t J(s)dB(s) + \int_0^t X^-(s)dM(s)$$

lo que implica que, si el rango de $X(t)$ es completo para todo t , entonces $\hat{B}(t)$ es un estimador insesgado de B(t), ya que la media de la martingala $\int_0^t X^-(s)dM(s)$ es cero [3].
(J(t) es un indicador de que la inversa de la matriz $\mathbf{X}$(t) existe, será 1 si existe y 0 sino.)

3.3.3. Hipótesis del modelo

Las hipótesis de este modelo se basan en comprobar por un lado la significación de las covariables y por otra lado en determinar si estas covariables varían a lo largo del tiempo.

Empezaremos explicando la hipótesis de significación:

$$H_{01} : \beta_p(t) \equiv 0 \qquad \text{para todo } t$$

que también puede formularse para los coeficientes de regresión acumulados como:

$$H_{01} : B_p(t) \equiv 0 \qquad \text{para todo } t$$

Rechazar la hipótesis nula para una covariable, significa que hay evidencias suficientes para decir que esta covariable es significativa, por lo que se quedará como parte del modelo.

Existen dos estadísticos para verificar esta hipótesis:

$$\tilde{T}_{1S} = \sup_{t\,\epsilon\,[0,\tau]} \mid \hat{B}_p(t) \mid$$

$$\tilde{T}_{1I} = \sup_{t,s\,\epsilon[0,\tau]} \mid \hat{B}_p(s) - \hat{B}_p(t) \mid$$

Respecto a la distribución de los estadísticos $\tilde{T}_{1S}$ y $\tilde{T}_{1I}$ cabe mencionar que bajo el modelo aditivo de Aalen y la Condición 1 [4] se establece que $U^{(n)} = n^{1/2}(\hat{B}(t) - B)$ converge en distribución hacia una martingala Gaussiana U, con función de varianza $\phi(t)$ (dependiente del tiempo).

El estadístico $\tilde{T}_{1I}$ es más eficiente para detectar desviaciones de $\beta_p(t)$, ya que $\tilde{T}_{1S}$ tiene poca potencia si $\beta_p(t)$ sólo difiere sustancialmente de 0 hacia el final del periodo $[0, \tau]$.

---

[3] Sea N_t un proceso contador con un compensador continuo A_t , tal que $M_t = N_t - A_t$ es una martingala de media cero. Entonces si H_t es cualquier proceso predecible definido en la misma filtración, el proceso $\int_0^t H_s dM_s$ es una martingala de media cero.

[4] Condición 1:

(a) $\sup_{t\,\epsilon[0,\tau]} E(Y_i(t)W_i^2(t)X_{ij}(t)Xik(t)X_{il}(t)) < \infty$ para todo$j, k, l = 1, ..., p$

(b) $r_2(t) = E(Y_i(t)W_i(t)X_i^{\otimes 2}(t))$es no singular para todo$t \in [0, \tau]$.

La segunda hipótesis del modelo sirve para determinar si los efectos de las covariables son constantes o varían a lo largo del tiempo.

$$H_{02} : \beta_p(t) \equiv \gamma, \qquad \text{para todo } t$$

que también puede expresarse para los coeficientes de regresión acumulados como:

$$H_{02} : B_p(t) \equiv \gamma t, \qquad \text{para todo } t$$

La hipótesis nula dice que el coeficiente de la covariable es constante en el tiempo, por lo que rechazar esta hipótesis supone asumir que el coeficiente de esa covariable es una función dependiente del tiempo. La comprobación de esta hipótesis puede hacerse mediante el test basado en Kolmogorov-Smirnov o mediante el test alternativo de Cramer von Mises.

Los estadísticos de los tests basados en Kolmogorov-Smirnov ($\tilde{T}_{2KS}$) y en Cramer von Mises ($\tilde{T}_{2CM}$) son los siguientes:

$$\tilde{T}_{2KS} = n^{1/2} \sup_{t \,\epsilon[0,\tau]} \mid \hat{B}_p(t) - \frac{t}{\tau}\hat{B}_p(\tau) \mid$$

y

$$\tilde{T}_{2CM} = n \int_0^\tau (\hat{B}_p(t) - \frac{t}{\tau}\hat{B}_p(\tau))^2 dt,$$

donde $\hat{B}_p(\tau)/\tau$ es una estimación de la constante subyacente bajo la hipótesis nula. El proceso del test básico en este contexto es

$$n^{1/2}(\hat{B}_p(t) - \frac{t}{\tau}\hat{B}_p(\tau))$$

para todo $t \in [0, \tau]$, que bajo la hipótesis nula tiene distribución asintótica.

3.3.4. Modelo de Aalen semiparamétrico

Cuando en el modelo de Aalen existen tanto covariables dependientes del tiempo como covariables que no lo son existe la posibilidad de hacer un modelo de Aalen semiparamétrico. En este modelo las covariables no dependientes del tiempo se considerán como la parte paramétrica. A continuación se muestra la expresión de la función de riesgo de este modelo:

$$\lambda(t|\boldsymbol{X}(t), \boldsymbol{Z}) = I(t) \left\{ \boldsymbol{X}(t)^T \boldsymbol{\beta}(t) + \boldsymbol{Z}(t)^T \boldsymbol{\gamma} \right\}$$

donde $\boldsymbol{\beta}(t)$ es un vector de funciones integrables de dimensión p, $\boldsymbol{\gamma}$ es un vector de dimensión q, $\boldsymbol{X}(t)$ y $\boldsymbol{Z}(t)$ son vectores de covariables de dimensión p y q respectivamente e $I(t)$ es un indicador de riesgo que toma valor 0 o 1.

Estimación de los parámetros

En este caso la descomposición de la martingala del proceso contador es

$$dN(t) = \lambda(t)dt + dM(t) = X(t)dB(t) + Z(t)\gamma dt + dM(t)$$

donde $X(t) = (I_1(t)X_1(t), ..., I_n(t)X_n(t))^T$ y $Z(t) = (I_1(t)Z_1(t), ..., I_n(t)Z_n(t))^T$.

Como los incrementos de la martingala no están correlacionados y su media es cero, $dB(t)$ y γ pueden estimarse mediante técnicas de mínimos cuadrados y de esta forma llegar a la siguiente expresión del estimador de $B(t)$:

$$d\hat{B}(t) = X^{-}(t)\left\{dN(t) - Z(t)\gamma dt\right\}$$

que se puede escribir de la forma

$$\hat{B}(t) = \int_0^t X^{-}(t)\left(dN(s) - Z(s)\hat{\gamma}(t)ds\right)$$

3.4. Implementación en R

A continuación se explican aquellos paquetes de R que se han utilizado en este trabajo y que no se estudian en el máster.

3.4.1. Paquete *muhaz* para estimar la función de riesgo

El paquete `muhaz` [19] se utiliza para calcular la estimación de la función de riesgo para datos censurados.

Para la estimación de la función de riesgo se ha utilizado la función `kphaz.fit` que calcula la estimación del riesgo mediante Kaplan-Meier.

```
kphaz.fit(time,status,strata,q=1,method="nelson")
```

donde `time` representa el tiempo, que en nuestro caso será el tiempo de supervivencia, `status` es la variable de censura, `strata` es opcional por si se quiere hacer la función de riesgo estratificada, `q` número de tiempos de fallo (por defecto es igual a 1) y `method` es el método para hacer la estimación que por defecto es el método "nelson".

Al hacer el gráfico de la función de riesgo mediante la función `kphaz.plot`, se encontró una limitación, ya que la función define los valores máximos de los ejes, por lo que para poder cambiarlos hubo que modificar la función (se muestra en el Apéndice B).

3.4.2. Paquete *timereg* para estimar el modelo aditivo de Aalen

El paquete `timereg` [25] se utiliza para hacer modelos flexibles de regresión con datos de supervivencia. En este trabajo se ha utilizado para ajustar el modelo aditivo de Aalen. Para ello se utiliza la función `aalen`:

```
aalen(formula,data=sys.parent(),start.time=0,max.time=NULL,robust=1,id=NULL,clusters=NULL
,residuals=0,n.sim=1000,weighted.test=0,covariance=0,resample.iid=0,deltaweight=1,silent=1
,weights=NULL,max.clust=1000, gamma=NULL, offsets=0)
```

Para más detalle puede consultarse la ayuda de la función en R (`help(aalen)`).

En este trabajo la función de Aalen ejecutada ha sido la siguiente:

```
aalen<-aalen(Surv(msurv, censura)~sex+tipo_aguda+ediag_categ+adiag_categ,aguda,max.time=60)
summary(aalen)
```

donde `Surv(msurv, censura)` crea el objeto supervivenvia, `~sex+tipo_aguda+ediag_categ+adiag_categ` son las covariables del modelo (variables independientes), `aguda` es la base de datos y `max.time=60` es el tiempo máximo de supevivencia.

Para el modelo aditivo de Aalen semi-paramétrico se usa el mismo comando `aalen`, pero hay que indicar en la función cuales son las covariables que se quiere que sean constantes. A continuación se muestra el modelo semiparamétrico ajustado en este trabajo:

```
aalen.semi<-aalen(Surv(msurv, censura)~const(sex)+tipo_aguda+ediag_categ+const(adiag_categ)
,aguda,max.time=60)
```

Capítulo 4

Resultados

4.1. Descripción de los datos

En los últimos 39 años, entre el año 1973 y 2012, ha habido 50572 pacientes diagnosticados de leucemia aguda en Estados Unidos. De ellos el 84.36 % son blancos[1](en 2012 la población de blancos en Estados Unidos era del 72.4 % [4]), el 83 % ha tenido un único tumor primario y solo un 7.8 % ha recibido tratamiento con radioterapia además de la quimioterapia (información completa en la Tabla A.1 del Apéndice A). La media de edad en el momento del diagnóstico es de 52.47 años. Y de los 50572 diagnosticados han fallecido 29025 (57.4 %), siendo la mediana 13.5 meses.

Como puede verse en la Tabla 4.1 la tasa de supervivencia global cada diez años ha mejorado, pasando de un 32.91 % entre los años 1973-1980 a un 49.25 % entre los años 2000-2012.

Si se analiza la supervivencia según el tipo de leucemia aguda se ve que existe una gran diferencia entre los distintos tipos. Los pacientes diagnosticados de leucemia linfocítica tienen una supervivencia a los 5 años del 67 %, mientras que los diagnosticados por cualquiera de los otros 3 tipos tienen una supervivencia menor al 40 %. Estos resultados podrían deberse a que la leucemia linfocítica es la más común en niños menores de 14 años, que a su vez son los pacientes que más sobreviven a la enfermedad.

[1]Traducción literal de la fuente, que utiliza white y black para referirse a la raza.

Variable	Categoría	N (%)	Mediana (meses)	IC (95 %)	Supervivencia a los 5 años
Género	Hombre	27661 (54.69)	13.5	(13.5, 14.5)	42.5 %
	Mujer	22911 (45.31)	13.5	(12.5, 13.5)	42.76 %
Edad al diagnóstico	<14 años	8289 (16.39)	NA	NA	80 %
	[14,45)	8856 (17.51)	NA	NA	51.11 %
	[45,65)	10941 (21.64)	12.5	(11.5, 12.5)	37.34 %
	$\geq$ 65	22483 (44.46)	3.5	(3.5, 3.5)	28 %
Tipo de leucemia aguda	Linfocítica	12492 (24.70)	NA	NA	67 %
	Mieloide	31863 (63.01)	7.5	(7.5, 8.5)	34.46 %
	Monocítica	2198 (4.35)	6.5	(5.5, 7.5)	30.53 %
	Otras	4010 (7.94)	4.5	(4.5, 5.5)	38.05 %
Año del diagnóstico	[1973,1980)	6148 (12.16)	9.5	(8.5, 9.5)	32.91 %
	[1980-1990)	11053 (21.86)	11.5	(11.5, 12.5)	38.14 %
	[1990-2000)	13279 (26.26)	12.5	(11.5, 13.5)	40.76 %
	[2000-2012]	20092 (39.73)	18.5	(17.5, 19.5)	49.25 %

Tabla 4.1: Mediana del tiempo de supervivencia y probabilidad de sobrevivir a los 5 años.

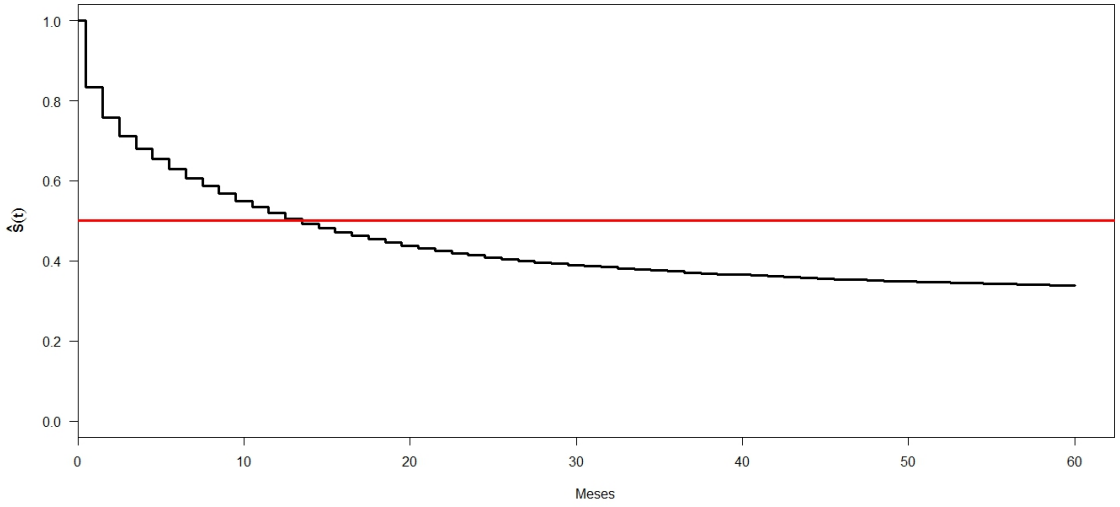

Figura 4.1: Función de supervivencia estimada.

La Figura 4.1 muestra la curva de supervivencia hasta los 5 años, y puede observarse que el 50 % de los pacientes fallecen antes de los 15 meses. También hay que destacar que en los primeros meses después del diagnóstico es donde se ven los saltos mas grandes, lo que quiere decir que son los meses donde más gente muere. Después la curva se va alisando hasta ser casi constante.

Esto se ve reflejado también en la función de riesgo, ya que se puede concluir que el riesgo más alto está en los primeros meses desde el diagnóstico. En la Figura 4.2 se puede observar como la función de riesgo después de los 3 meses del diagnóstico desciende a la mitad, pasando de un 0.1 a 0.05. Después de estos 3 meses el riesgo va disminuyendo de forma más constante hasta ser casi cero a los 5 años.

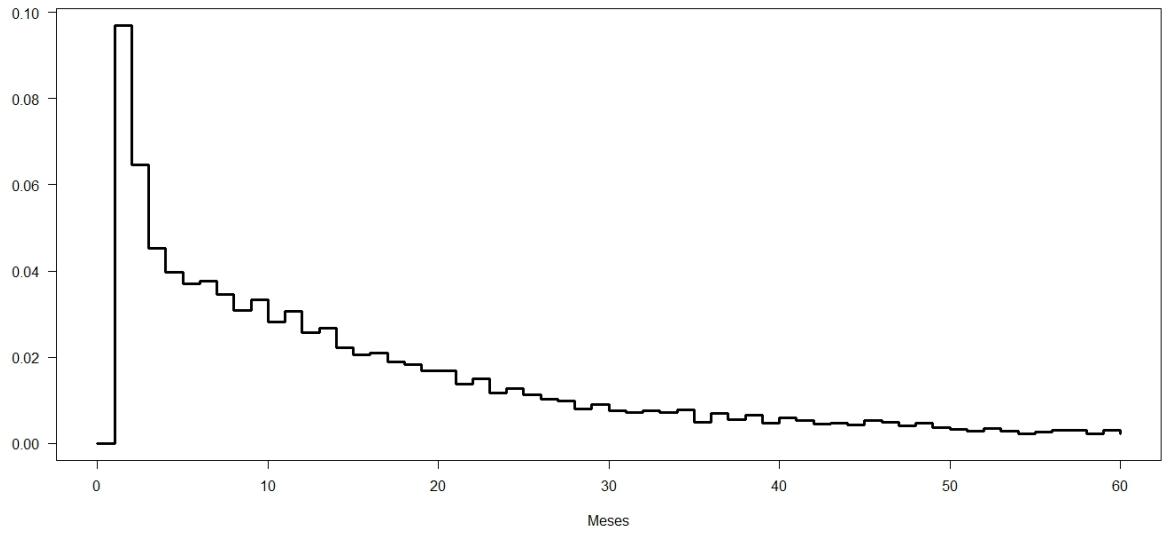

Figura 4.2: Función de riesgo de pacientes de leucemia aguda en los 5 años siguientes al diagnóstico.

Si se examina la función de riesgo en función de las categorías de cada covariable, se concluye que según el género y año del diagnóstico, cada categoría tiene el mismo comportamiento que la función de riesgo general. Es decir, en los primero meses el riesgo desciende rápidamente y en los siguiente meses decrece más constantemente hasta llegar casi a cero (Figuras A.1 y A.2 del Apéndice A).

Este mismo resultado se obtiene con la variable tipo de leucemia aguda, salvo por una excepción que es la leucemia linfocítica aguda que tiene una función de riesgo casi constante e igual a cero desde el inicio (Figura A.3).

Sin embargo al analizar la función de riesgo según la edad en el momento del diagnóstico, como se puede apreciar en la Figura 4.3, los comportamientos son distintos. Por un lado están los menores de 14 años, que tienen como puede verse en su curva de la función de riesgo, un riesgo casi cero ya que la curva es casi plana e igual a cero. Por otro lado hay dos grupos con el mismo comportamiento, que son los pacientes con edades comprendidas entre 14 y 45 años y el grupo de pacientes que tiene entre 45 y 65 años. Estos dos grupos tienen un riesgo más alto que los menores de 14 años, pero aún así tienen un riesgo bajo siendo constante a lo largo del tiempo.

Por último están los mayores de 65 años que tienen una curva de la función de riesgo igual a la curva de la función de riesgo general, siendo el riesgo más alto en los primeros meses y decreciendo hasta hacerse casi cero en los últimos meses.

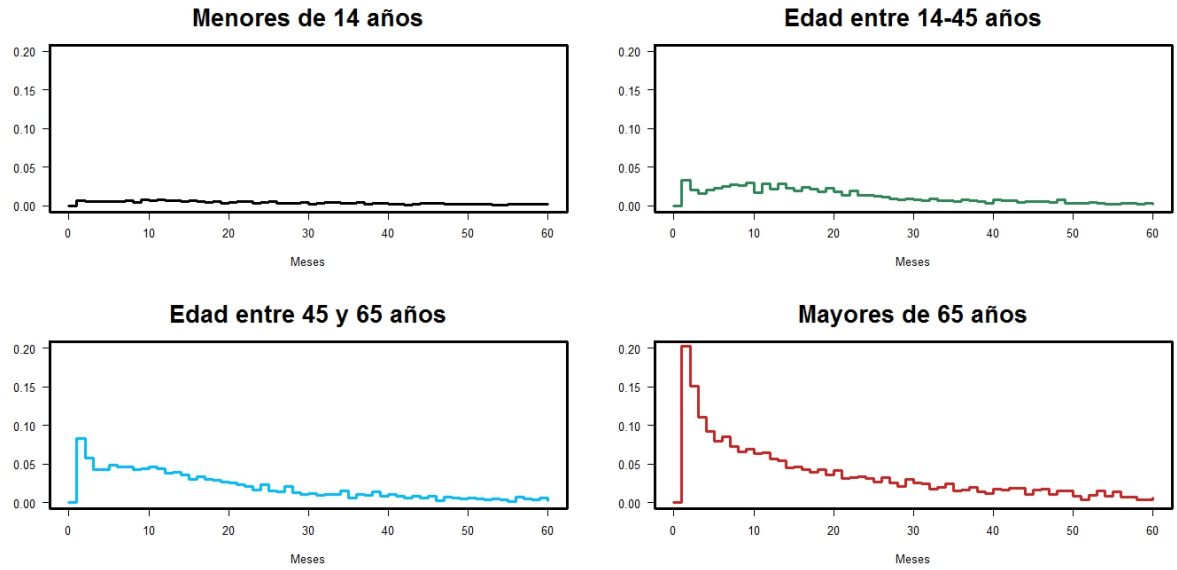

Figura 4.3: Función de riesgo según la edad del paciente en el momento del diagnóstico. De izquierda a derecha y de arriba abajo son, menores de 14 años, edad entre 14 y 65 años, entre 45 y 65 años y mayores de 65 años.

4.2. Resultados según un modelo de Cox

A continuación se muestran los resultados del análisis de la supervivencia en la leucemia aguda según un modelo de Cox. La expresión del modelo es el siguiente:

$$\lambda(t|\mathbf{X}) = \lambda_0(t)\ \exp\{\beta'\ \mathbf{X}\} = \lambda_0(t)\ \exp\{\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p\}$$

Las covariables son género (Ref: Hombre), tipo de leucemia aguda (Ref: Leucemia linfocítica aguda), edad en el diagnóstico (Ref: Menores de 14 años) y año del diagnóstico (Ref: 1973-1980). Así con las covariables dummy creadas se incluyen un total de 10 covariables. Por lo tanto la expresión del modelo para este estudio es:

$$\lambda(t|\mathbf{X}) = \lambda_0(t) \exp\left\{\sum_{j=1}^{10} \beta_j X_j\right\}$$

En la Tabla 4.2 se puede observar que todas las variables excepto la de género son estadísticamente significativas utilizando $\alpha = 0.05$ como nivel de significación, pero aún así se decidió dejar esta variable en el modelo ya que es una variable importante en epidemiología.

Si se observa la supervivencia según el tipo de leucemia aguda se puede ver que el riesgo de un paciente con leucemia linfocítica aguda es menor que el de los pacientes con igual perfil de covariables pero con otro tipo de leucemia aguda. Por ejemplo los pacientes diagnosticados de leucemia mieloide aguda tienen un riesgo instantáneo de morir un 47 % más alto que los diagnosticados de leucemia linfocítica aguda. Los del tipo monocítica tienen 82 % más de riesgo que los pacientes de linfocítica y el resto de leucemias agudas un 58 % más de riesgo.

Si se examinan los coeficientes de la variable edad al diagnóstico se ve que van aumentando a medida que aumenta la edad, siempre mayores que uno, lo que indica que cuanto mayor sea la edad en el momento del diagnóstico mayor será el riesgo de morir en un plazo de 5 años. Por ejemplo los pacientes con edades entre los 45 y 65 años tienen un riesgo instantáneo de morir que es 4.73 veces mayor que los menores de 14 años.

Por último sobre el año del diagnóstico se puede concluir que cuanto más tarde han sido diagnosticados menor es el riesgo de morir, un resultado esperado ya que como se ha comentado al inicio del capítulo, la supervivencia global ha mejorado a lo largo de los años gracias a la investigación y aparición de nuevos fármacos.

Variable		**Coeficiente**	**HR**	**IC(95 %)**	**P-valor**
Género	Mujer	-0.013	0.987	(0.965, 1.010)	0.277
Tipo de leucemia aguda	Mieloide	0.391	1.478	(1.421, 1.536)	$< 2e^{-16}$
	Monocítica	0.601	1.823	(1.714, 1.939)	$< 2e^{-16}$
	Otra	0.458	1.581	(1.498, 1.668)	$< 2e^{-16}$
Edad al diagnóstico	$[14 - 45)$	1.031	2.805	(2.642, 2.978)	$< 2e^{-16}$
	$[45 - 65)$	1.554	4.732	(4.456, 5.025)	$< 2e^{-16}$
	≥ 65	2.225	9.258	(8.731, 9.816)	$< 2e^{-16}$
Año del diagnóstico	$[1980, 1990)$	-0.179	0.836	(0.804, 0.869)	$< 2e^{-16}$
	$[1990, 2000)$	-0.309	0.735	(0.707, 0.763)	$< 2e^{-16}$
	$[2000, 2012]$	-0.549	0.577	(0.557, 0.599)	$< 2e^{-16}$

Tabla 4.2: Resultados del modelo de Cox donde HR es el hazard ratio. Las categorías de referencia son hombre, con leucemia linfocítica aguda, menor de 14 años y diagnosticado entre los años 1973 y 1980.

Validación del modelo

A continuación se presenta la validación del modelo para comprobar si las covariables cumplen la premisa de riesgos proporcionales. Esta validación se hará mediante los residuos de Schoenfeld.

En la Figura 4.4 que se presenta a continuación se muestran los gráficos de los residuos de Schoenfeld para cada covariable. Como se ha explicado en el Capítulo 3, para que haya proporcionalidad para una covariable, la curva del estimador de esa covariable en función del tiempo debe ser una línea recta con pendiente cero. En los gráficos puede verse que las covariables que

cumplen esta condición son género y año del diagnóstico. En cambio, se puede ver como las curvas de los estimadores de las covariables tipo de leucemia aguda y edad en el diagnóstico no se ajustan a una línea recta con pendiente cero, por lo que estas dos covariables no cumplen los riesgos proporcionales.

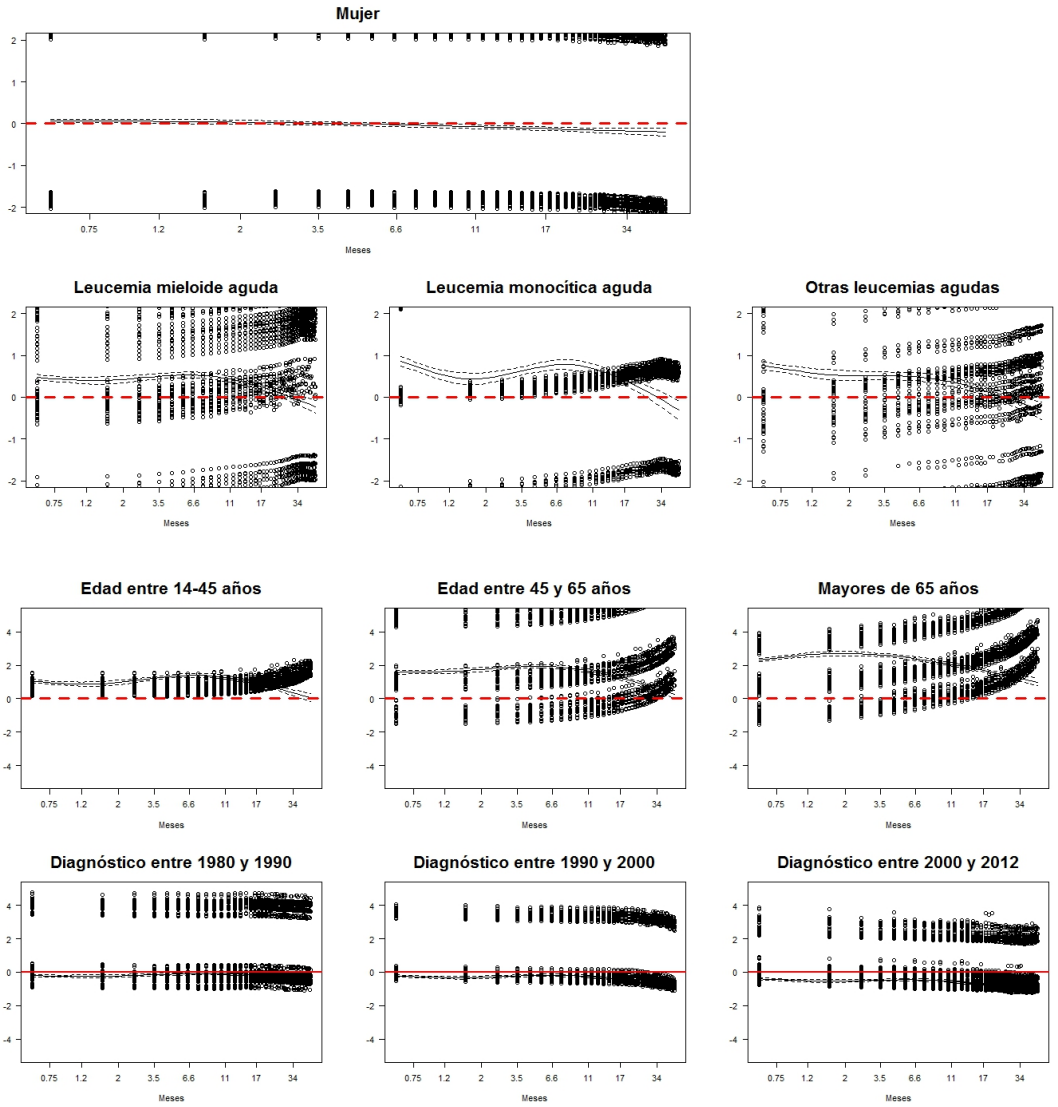

Figura 4.4: Gráficos de los residuos de Schoenfeld de las covariables género y tipo de leucemia aguda. La curva del estimador de cada covariable está representada por una línea continua.

A continuación se comprueba la hipótesis de que los riesgos de dos individuos con covariables diferentes son proporcionales analíticamente. El resultado se puede ver en la Tabla 4.3 donde se muestran los resultados de los residuos de Schoenfeld. Según estos resultados se podría concluir que los efectos de todas las covariables no son constantes (p-valores <0.05, excepto la categoría

[1980-1990) que tiene un p-valor > 0.05), pero no obstante los resultados que se tendrán en cuenta son los obtenidos mediante los gráficos, ya que estos valores de p tan pequeños pueden deberse al gran número de datos que contiene la base de datos.

Variable		**rho**	χ^2	**p-valor**
Género	Mujer	-0.039	44.97	$2.00\ e^{-11}$
Tipo de leucemia aguda	Mieloide	-0.031	31.25	$2.27e^{-08}$
	Monocítica	-0.036	39.41	$3.43e^{-10}$
	Otra	-0.056	93.58	$< e^{-16}$
Edad al diagnóstico	$[14-45)$	-0.014	5.97	$1.45e^{-02}$
	$[45-65)$	-0.032	32.65	$1.10e^{-08}$
	≥ 65	-0.056	98.65	$< e^{-16}$
Año del diagnóstico	$[1980, 1990)$	0.006	1.08	$2.98e^{-01}$
	$[1990, 2000)$	-0.028	22.58	$2.02e^{-06}$
	$[2000, 2012]$	-0.049	68.26	$1.11e^{-16}$
GLOBAL		NA	672.97	$< e^{-16}$

Tabla 4.3: Validación del modelo de Cox (rho es el coeficiente de correlación entre el tiempo de supervivencia transformado y los residuos de Schoenfeld escalados).

4.2.1. Modelo de Cox estratificado

Como se explica en el Capítulo 3, cuando alguna variable no cumple los riesgos proporcionales existe la posibilidad de ajustar un modelo Cox estratificado por esa covariable.

Como se ha visto en el apartado anterior, algunas covariables del modelo no cumplen la condición de proporcionalidad. Por este motivo se estratificará por la covariable tipo de leucemia aguda. Esta covariable ha sido elegida entre todas las demás porque además de no cumplir los riesgos proporcionales es razonable pensar que cada tipo de leucemia actúa como una enfermedad distinta al resto y por ello tener características y tratamientos distintos.

En la Tabla 4.4 se presentan los resultados del modelo de Cox estratificado por la covariable tipo de leucemia aguda.

Variable		Coeficiente	HR	IC(95 %)	P-valor
Género	Mujer	-0.012	0.988	(0.966, 1.012)	0.32
Edad al diagnóstico	$[14-45)$	1.090	2.975	(2.8, 3.160)	$< e^{-16}$
	$[45-65)$	1.619	5.046	(4.746, 5.364)	$< e^{-16}$
	≥ 65	2.276	9.739	(9.176, 10.338)	$< e^{-16}$
Año del diagnóstico	$[1980, 1990)$	-0.18	0.836	(0.804, 0.869)	$< e^{-16}$
	$[1990, 2000)$	-0.309	0.734	(0.707, 0.763)	$< e^{-16}$
	$[2000, 2012]$	-0.549	0.577	(0.557, 0.599)	$< e^{-16}$

Tabla 4.4: Resultados del modelo de Cox estratificado por tipo de leucemia aguda.

Igual que ocurría para el modelo general, la variable género no es estadísticamente significativa (p-valor > 0.05), por lo que no se puede afirmar que haya diferencias en la supervivencia entre hombres y mujeres.

Para el resto de las covariables los resultados tampoco difieren mucho del modelo general, extrayendo las mismas conclusiones. Cuanto mayor es la edad del paciente en el momento del diagnóstico mayor es el riesgo de morir. Por ejemplo los pacientes diagnosticados a los 65 años o más, tienen casi 10 veces más de riesgo instantáneo de morir que los menores de 14 años, y los pacientes de entre 14 y 45 años tienen un riesgo casi tres veces mayor.

Respecto al año del diagnóstico, la supervivencia mejora cuanto mas tarde haya sido, siendo la mitad en los diagnosticados entre los años 2000 y 2012 que a los que se les diagnosticó entre 1973 y 1980.

Si se comparan estos resultados con los obtenidos con el modelo de Cox sin estratificar (Tabla 4.2) mediante los intervalos de confianza del hazard ratio se observa que los valores son similares, siendo los valores de los intervalos de confianza del modelo estratificado valores más altos.

Validación del modelo estratificado

A continuación se hace la validación del modelo de Cox estratificado mediante los residuos de Schoenfeld. En la Figura 4.5 se muestran los gráficos de estos residuos para cada covariable. En ellos se puede ver que las covariables género y año del diagnóstico cumplen los riesgos proporcionales mientras que la covariable edad en el diagnóstico continua sin cumplirlos.

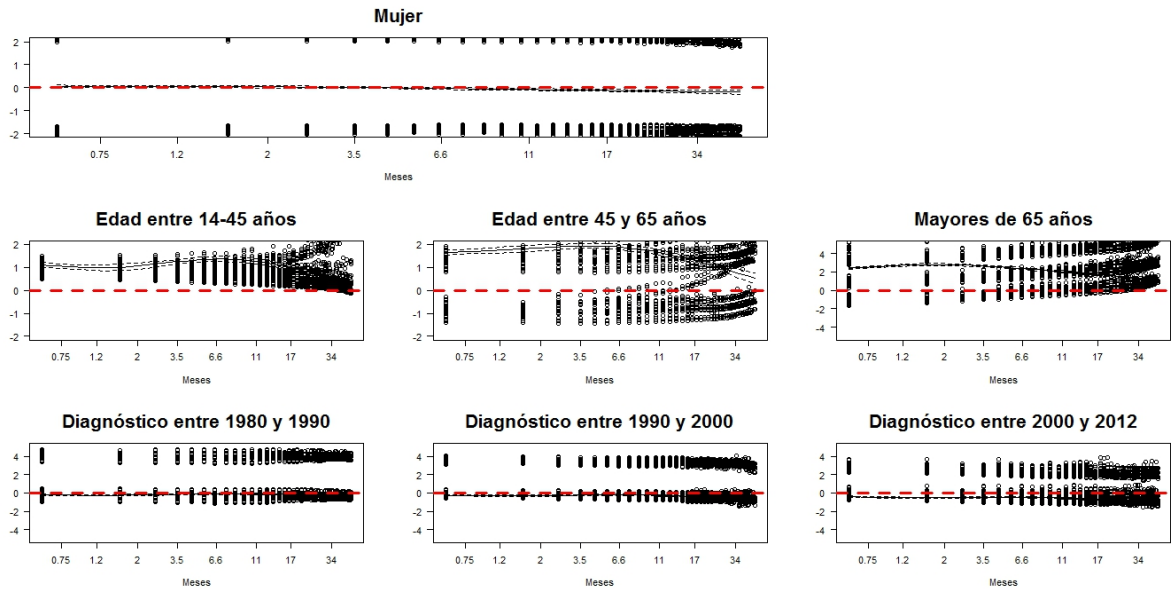

Figura 4.5: Residuos de Schoenfeld del modelo de Cox estratificado por tipo de leucemia aguda para las covariables género, edad en el diagnóstico y año del diagnóstico.

En la Tabla 4.5 se muestran los resultados del análisis de los residuos de Schoenfeld. Igual que ocurría con el modelo de Cox sin estratificar según los p-valores se podría concluir que los efectos de todas las covariables (exceptuando la categoría $[1980, 1990)$ de la covariable año del diagnóstico) no son constantes (p-valores < 0.05, excepto la categoría [1980-1990) que tiene un p-valor > 0.05), pero no obstante los resultados que se tendrán en cuenta son los obtenidos mediante los gráficos, ya que estos valores de p tan pequeños pueden deberse al gran número de datos que contiene la base de datos.

Variable		**rho**	χ^2	**p-valor**
Género	Mujer	-0.039	43.814	$3.61e^{-11}$
Edad al diagnóstico	$[14-45)$	-0.013	5.606	$1.79e^{-02}$
	$[45-65)$	-0.031	31.463	$2.03e^{-08}$
	≥ 65	-0.056	100.604	$< e^{-16}$
Año del diagnóstico	$[1980, 1990)$	0.006	0.946	$3.31e^{-01}$
	$[1990, 2000)$	-0.028	22.374	$2.24e^{-06}$
	$[2000, 2012]$	-0.048	67.717	$2.22e^{-16}$
GLOBAL		NA	429.122	$< e^{-16}$

Tabla 4.5: Validación del modelo de Cox estratificado (rho es el coeficiente de correlación entre el tiempo de supervivencia transformado y los residuos de Schoenfeld escalados).

Por todo lo mencionado anteriormente, se puede concluir que estratificar el modelo mediante la variable tipo de leucemia aguda no soluciona el problema de proporcionalidad de los datos que surgía en el modelo de Cox, por lo que entre los dos modelos se escoge el modelo de Cox sin estratificar ya que así se tienen más covariables con las que estimar y comparar la supervivencia.

4.3. Resultados según el modelo aditivo de Aalen

El ajuste del modelo aditivo de Aalen ha proporcionado los gráficos de las estimaciones de la función de regresión acumulada de cada covariable que se muestran a continuación en la Figura 4.6.

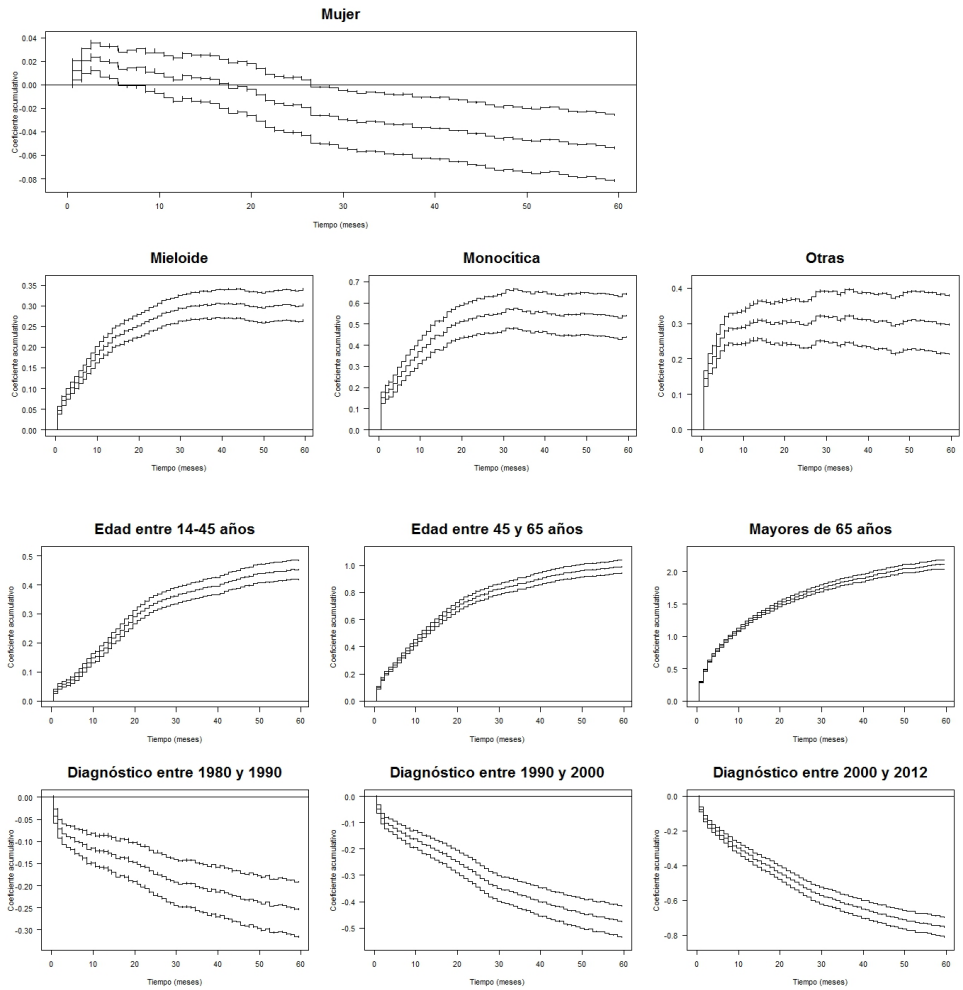

Figura 4.6: Gráficos de las estimaciones de la función de regresión acumulada del modelo de Aalen de las covariables género, tipo de leucemia aguda, edad en el diagnóstico y año del diagnóstico.

En los gráficos se puede ver que los coeficientes de las estimaciones de la función de regresión

acumulada de las covariables género y año del diagnóstico son líneas rectas, lo que indica que los efectos de estas covariables no varían en el tiempo. En cambio en los gráficos de las covariables tipo de leucemia aguda y edad en el diagnóstico se ve que los coeficientes de las estimaciones de la función de regresión acumulada son curvas en lugar de líneas rectas, lo que indica que los efectos de estas covariables varían en el tiempo. Además en el caso de estas dos covariables se observa que tienen un fuerte efecto en los primeros 20 meses desde el diagnóstico. Después de estos 20 meses este efecto se disipa hasta convertirse las curvas en líneas rectas. Esto puede deberse a que el riesgo de morir en los pacientes baja casi a cero después de los 20 meses como puede verse en la Figura 4.2.

Respecto a las pendientes de las curvas de los coeficientes las covariables tipo de leucemia aguda y edad en el diagnóstico tienen pendiente positiva y los valores están por encima de cero lo que indica que el riesgo es mayor que en la categoría de referencia, que en estos dos casos son leucemia linfocítica aguda y menores de 14 años. Las covariables género y año del diagnóstico tienen pendiente negativa y los valores están por debajo de cero lo que indica que el riesgo es menor que en la categoría de referencia, que es hombre y año del diagnóstico entre el año 1973 y 1980. Respecto a la covariable género hay que tener en cuenta que los valores del eje Y que aparecen en el gráfico son muy pequeños.
A continuación en la Tabla 4.6 se muestran los p-valores del test de significación y de los tests de Kolmogorov-Smirnov y de Cramer von Mises.

Variable		**Test de significación**	**Test Kolmogorov-Smirnov**	**Test Cramer von Mises**
Género	Mujer	$< e^{-16}$	0.003	0.022
Tipo de leucemia aguda	Mieloide	$< e^{-16}$	$< e^{-16}$	$< e^{-16}$
	Monocítica	$< e^{-16}$	$< e^{-16}$	$< e^{-16}$
	Otra	$< e^{-16}$	$< e^{-16}$	$< e^{-16}$
Edad al diagnóstico	$[14-45)$	$< e^{-16}$	$< e^{-16}$	$< e^{-16}$
	$[45-65)$	$< e^{-16}$	$< e^{-16}$	$< e^{-16}$
	≥ 65	$< e^{-16}$	$< e^{-16}$	$< e^{-16}$
Año del diagnóstico	$[1980, 1990)$	$< e^{-16}$	$< e^{-16}$	$< e^{-16}$
	$[1990, 2000)$	$< e^{-16}$	$< e^{-16}$	$< e^{-16}$
	$[2000, 2012]$	$< e^{-16}$	$< e^{-16}$	$< e^{-16}$

Tabla 4.6: Efectos aditivos de las covariables bajo el modelo aditivo de Aalen.

Según el test de significación todas las covariables son significativas ya que todos los p-valores son menores de 0.05.

Los tests de Kolmogorov-Smirnov y Cramer von Mises se utilizan para determinar si los efectos de las covariables son constantes o varían a lo largo del tiempo. Los dos test utilizan como hipótesis nula que los coeficientes de las covariables son constantes y no varían en el tiempo. Por lo tanto según los resultados que se observan en la Tabla 4.6 las covariables de este modelo

varían en el tiempo, ya que según los p-valores que se muestran (todos ellos menores que 0.05) se debe rechazar la hipótesis nula de que los efectos de las covariables son constantes en el tiempo. Sin embargo, igual que se ha hecho al ajustar el modelo de Cox las conclusiones obtenidas gráficamente serán las que se tendrán en cuenta, ya que estos p-valores tan pequeños pueden ser causa del gran número de datos que contiene la base.

4.3.1. Modelo de Aalen semiparamétrico

Como se explica en el Capítulo 3, cuando existen algunas covariables que no varían en el tiempo el modelo aditivo de Aalen se puede flexibilizar mediante un modelo semiparamétrico donde se fijan las covariables con efectos constantes. En este caso se considerará que las covariables género y año del diagnóstico no varían en el tiempo.

Al hacer el modelo semiparamétrico es posible obtener el valor de los coeficientes de las covariables que se presuponen constantes sus efectos, como se muestra en la Tabla 4.7. Se puede observar que la variable género no es estadísticamente significativa pero que el año del diagnóstico si. Además igual que ocurría en el modelo de Cox, cuanto mas tarde haya sido el diagnóstico menor es el riesgo de morir.

Variable		**Coeficiente**	**HR**	**ES**	**P-valor**
Género	Mujer	0.00	1	0.00	0.28
Año del diagnóstico	$[1980, 1990)$	-0.007	0.993	0.001	$< e^{-16}$
	$[1990, 2000)$	-0.011	0.989	0.001	$< e^{-16}$
	$[2000, 2012]$	-0.018	0.982	0.001	$< e^{-16}$

Tabla 4.7: Resultados de las covariales paramétricas del modelo de Aalen semiparamétrico, donde HR es el harzard ratio y ES el error estándar.

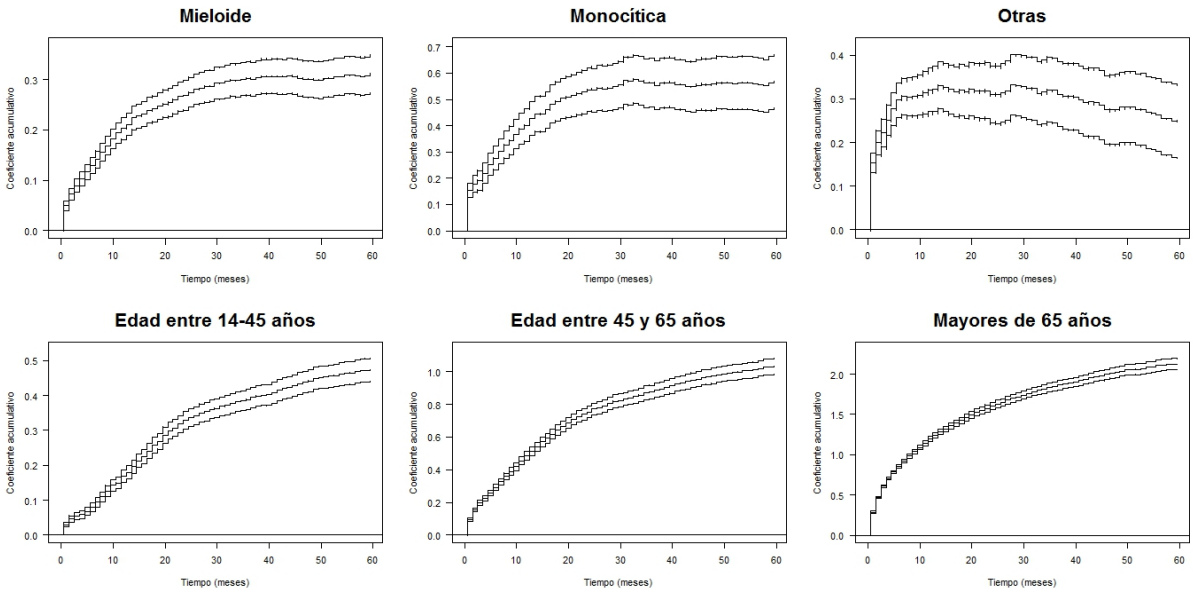

Figura 4.7: Gráficos de las estimaciónes de la función de regresión acumulada del modelo semiparamétrico de Aalen de las covariables tipo de leucemia aguda y edad en el diagnóstico.

En la Figura 4.7 se muestran las estimaciones de las funciones de regresión acumulada de las variables que varían en el tiempo, que son tipo de leucemia aguda y edad en el diagnóstico. Como se preveía, los coeficientes de estas covariables no son líneas rectas sino que curvas ascendentes.

Otro detalle que puede apreciarse en los gráficos es que las curvas a partir de cierto periodo de tiempo, alrededor de los 25 meses, se convierten en líneas rectas. Esto puede deberse a que como se ha visto en la curva de la función de riesgo (Figura 4.2), a partir de los 20 meses el riesgo es constante y casi cero.

Por todo lo explicado anteriormente, entre el modelo aditivo de Aalen y el modelo semiparamétrico de Aalen el modelo semiparamétrico con las covariables género y año del diagnóstico como constantes es el modelo preferible para estimar la supervivencia de leucemia aguda, ya que permite obtener el valor de la estimación de los coeficientes de las covariables que son constantes en el tiempo y ver gráficamente el comportamiento de las covariables dependientes de tiempo. Además las covariables siguen siendo significativas.

4.4. Comparación entre los dos modelos

A continuación en la Tabla 4.8 se muestran los resultados del modelo de Cox y del modelo aditivo de Aalen semiparamétrico.
Si se comparan los p-valores de cada covariable, se puede observar que las diferencias que hay entre los dos modelos son muy pequeñas, lo que se debe también al gran número de datos. El p-

Variable		β	β^*	**P-valor**	**P-valor***
Género	Mujer	-0.013	0.00	0.277	0.28
Tipo de leucemia aguda	Mieloide	0.391	-	$< 2e^{-16}$	$< 2e^{-16}$
	Monocítica	0.601	-	$< 2e^{-16}$	$< 2e^{-16}$
	Otra	0.458	-	$< 2e^{-16}$	$< 2e^{-16}$
Edad al diagnóstico	$[14 - 45)$	1.031	-	$< 2e^{-16}$	$< 2e^{-16}$
	$[45 - 65)$	1.554	-	$< 2e^{-16}$	$< 2e^{-16}$
	≥ 65	2.225	-	$< 2e^{-16}$	$< 2e^{-16}$
Año del diagnóstico	$[1980, 1990)$	-0.179	-0.007	$< 2e^{-16}$	$< 2e^{-16}$
	$[1990, 2000)$	-0.309	-0.011	$< 2e^{-16}$	$< 2e^{-16}$
	$[2000, 2012]$	-0.549	-0.018	$< 2e^{-16}$	$< 2e^{-16}$

Tabla 4.8: Comparación de los resultados del modelo de Cox y el modelo semiparamétrico de Aalen*.

valor de la covariable género es un poco menor en el modelo de Cox pero la significación de toda ellas es la misma, es decir, en los dos modelos género no es significativa y el resto de covariables sí.

Respecto a la validación de los modelos, en el modelo de Cox al no cumplirse la premisa de los riesgos proporcionales no ha podido ser validado. En cambio esto no es necesario en el modelo de Aalen ya que no tiene restricciones en las covariables.

Capítulo 5

Discusión y conclusiones

En este trabajo final de máster se ha presentado un modelo alternativo al modelo de Cox, el modelo aditivo de Aalen. Este modelo no se basa en la suposición de riesgos proporcionales; además, permite que los efectos de las covariables varíen a lo largo del tiempo.
Al utilizar la regresión de Cox es necesario verificar que se cumple la hipótesis de riesgos proporcionales, es decir, que el efecto de cada covariable es constante a lo largo del tiempo. Si no se cumple esta premisa para una covariable, el método comúnmente mas utilizado es el modelo de Cox estratificado, donde la función basal se reemplaza por tantas funciones basales como número de estratos haya. Pero a veces ocurre que aún así siguen sin cumplirse los riesgos proporcionales.

El modelo aditivo de Aalen permite que los coeficientes de las covariables sean funciones dependientes del tiempo. Además es posible construir un gráfico de la función de regresión acumulada de cada covariable donde se puede observar el comportamiento de cada una de ellas a lo largo del tiempo.

En este trabajo el objetivo ha sido comparar el modelo de Cox y el modelo aditivo de Aalen para datos de supervivencia por leucemia aguda, ya que por ser una patología de larga duración es lógico pensar que los efectos de las covariables no serán constantes a lo largo del tiempo. Para ello se ha utilizado una base de datos con la población diagnosticada de leucemia aguda en Estados Unidos desde el año 1973 hasta el año 2012 y las covariables seleccionadas han sido género, tipo de leucemia aguda, edad en el momento del diagnóstico y año en el que fue diagnosticado cada paciente.

Respecto a los resultados obtenidos, el signo de los coeficientes estimados a partir de los dos modelos son iguales, pero los coeficientes no pueden ser comparados ya que no miden exactamente lo mismo. Los coeficientes del modelo aditivo de Aalen están relacionados con las diferencias de las funciones de riesgo y los del modelo de Cox con la razón de las funciones de riesgo.

La conclusión bajo los dos modelos es que los factores de riesgo en la muerte por leucemia aguda son el tipo de leucemia aguda, la edad del paciente en el momento del diagnóstico y el año en que son diagnosticados. Además la leucemia linfocítica aguda es la que menor riesgo tiene, seguida por la leucemia mieloide aguda. Respecto a la edad en el diagnóstico se puede concluir que cuanto mayor sea el paciente, mayor será el riesgo de morir por leucemia aguda.Y por último

se ha visto que la supervivencia ha mejorado con los años.

Al ajustar el modelo aditivo de Aalen se obtiene que las covariables tipo de leucemia aguda y edad en el diagnóstico son covariables dependientes del tiempo, es decir, según en qué momento de la enfermedad se encuentre el paciente, el efecto de cada una de estas covariables en el riesgo será diferente. Por ejemplo, el riesgo de morir por un tipo concreto de leucemia varía en el tiempo en los primeros 10 meses desde el diagnóstico. Lo mismo ocurre con la edad en el diagnóstico en los primeros 20 meses.

Por lo tanto las conclusiones con los dos modelos son las mismas, pero el modelo aditivo de Aalen da más información cuando el efecto de las covariables varían en el tiempo. Entre los posibles inconvenientes de este modelo están que el criterio de Akaike (AIC)[1] y el criterio de información Bayesiano (BIC)[2] no se pueden utilizar para elegir el modelo, ya que la función de verosimilitud es difícil de especificar para este modelo. Además cuando se tiene un gran volumen de datos la ejecución del modelo de Aalen con el software R conlleva mucho tiempo de compilación, siendo en algunas ocasiones imposible de hacer.

Por todo lo mencionado anteriormente se puede concluir que si no se cumplen los riesgos proporcionales del modelo de Cox, el modelo aditivo de Aalen es una buena alternativa. Y si los riesgos proporcionales se cumplen los dos modelos son apropiados, ya que dan resultados similares respecto a las covariables. Por ello no deberían verse como alternativas, sino como métodos complementarios, ya que juntos dan una comprensión mas completa de los datos.

Como futuras investigaciones desde un punto metodológico sería interesante hacer este mismo estudio para los pacientes diagnosticados de leucemia crónica y ver si las conclusiones de la comparación de los dos modelos son las mismas que se han obtenido en este trabajo. Por otro lado, desde un aspecto clínico se puede plantear un estudio con los pacientes diagnosticados de leucemia aguda en España y comparar estos resultados con los obtenidos en este trabajo para los pacientes de Estados Unidos, y ver si hay diferencias en la supervivencia entre países. Y si las hubiera determinar si la causa es la diferencia entre tratamientos u otros factores.

[1] El AIC (criterio de información de Akaike) es una función de la verosimilitud maximizada (l) y el numero de parámetros estimables (K): $AIC = -2l + 2K$. Se escoge el modelo con el AIC mas pequeño.

[2] El BIC (criterio de información Bayesiano) es una función de la verosimilitud maximizada (l) y el numero de parámetros estimables (K): $BIC = -2l + 2 \log n$, donde n es el tamaño de la muestra. Se escoge el modelo con el BIC mas pequeño.

Bibliografía

[1] http://www.cancer.org/index, Abril 2016.

[2] http://www.cancer.gov, Abril 2016.

[3] http://seer.cancer.gov/, Abril 2016.

[4] http://www.census.gov, Abril 2016.

[5] http://www.aeal.es/, Abril 2016.

[6] http://www.who.int/en/, Abril 2016.

[7] Aalen, O., Borgan, O., and Gjessing, H. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.

[8] Abadi, A., Saadat, S., Yavari, P., Bajdik, C., and Jalili, P. Comparison of aalen's additive and cox proportional hazards models for breast cancer survival: analysis of population-based data from british columbia, canada. *Asian Pac J Cancer Prev 12* (2011), 3113–3116.

[9] Baldi, I., Ciccone, G., Ponti, A., Rosso, S., Zanetti, R., and Gregori, D. An application of the cox-aalen model for breast cancer survival. *Austrian journal of statistics 35*, 1 (2006), 77–88.

[10] Chandran, R., Gardiner, S. K., Smith, S. D., and Spurgeon, S. E. Improved survival in hairy cell leukaemia over three decades: a seer database analysis of prognostic factors. *British journal of haematology 163*, 3 (2013), 407–409.

[11] Gómez, G., and Delicado, P. *Curso de Inferencia y Decisión*. Universitat Politécnica de Catalunya, 2006.

[12] Gómez, G., Julià, O., and Langohr, K. *Análisis de supervivencia*. Universitat Politécnica de Catalunya and Universitat de Barcelona.

[13] Hernández, J. A., Land, K. J., and McKenna, R. W. Leukemias, myeloma, and other lymphoreticular neoplasms. *Cancer 75*, S1 (1995), 381–394.

[14] Klein, J. P., and Moeschberger, M. L. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2005.

[15] Llorca, J., and Delgado-Rodríguez, M. Análisis de supervivencia en presencia de riesgos competitivos: estimadores de la probabilidad de suceso. *Gaceta Sanitaria 18*, 5 (2004), 391–397.

[16] Luo, J. *SEER2R: reading and writing SEER*STAT data files*, 2012. R package version 1.0.

[17] Ma, H., Sun, H., and Sun, X. Survival improvement by decade of patients aged 0-14 years with acute lymphoblastic leukemia: a seer analysis. *Scientific reports 4* (2014).

[18] Martinussen, T., and Scheike, T. H. *Dynamic regression models for survival data*. Springer Science & Business Media, 2007.

[19] original by Kenneth Hess, S., and port by R. Gentleman, R. *muhaz: Hazard Function Estimation in Survival Analysis*, 2014. R package version 1.2.6.

[20] Páez, M. A., Borges, R., and Colmenares, G. Modelos de regresión aplicados a la banca comercial venezolana, período 1996-2004.

[21] Pereira, T. L., Colosimo, E. A., and Raposo, M. C. Modelo aditivo de aalen: uma aplicação para dados de sinusite em pacientes com aids. *Revista Colombiana de Estadística 30*, 1 (2007), 129–141.

[22] Posada, D., and Buckley, T. R. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic biology 53*, 5 (2004), 793–808.

[23] Scheike, T. H., and ZHANG, M.-J. An additive–multiplicative cox–aalen regression model. *Scandinavian Journal of Statistics 29*, 1 (2002), 75–88.

[24] Scheike, T. H., and Zhang, M.-J. Analyzing competing risk data using the r timereg package. *Journal of statistical software 38*, 2 (2011).

[25] Scheike, T. H., and Zhang, M.-J. *Analyzing Competing Risk Data Using the R timereg Package*, 2011.

Apéndice A

Anexo

A.1. Códigos oncológicos

- **CS**: Collaborative Stage data collection system (`https://cancerstaging.org/cstage/schema/Pages/version0204.aspx`).

- **ICD-O-2** : International Classification of diseases for Oncology, second edition (`http://seer.cancer.gov/tools/conversion/ICDO2-3manual.pdf`).

- **ICD-O-3** :International Classification oF diseases for Oncology, third edition (`http://apps.who.int/iris/bitstream/10665/96612/1/9789241548496_eng.pdf`).

- **ICD-9** : Conversion of Malignant Neoplasms by Topography and Morphology from the International Classification of Disease for Oncology, Second Edition (ICD-0-2) to International Classification of Diseases, 9th Revision (ICD-9) and the International Classification of Diseases, 9th Revision. (`http://www.cdc.gov/nchs/icd/icd9.htm`).

- **ICD-10** : Conversion of Malignant Neoplasms by Topography and Morphology from the International Classification of Disease for Oncology, Second Edition (ICD-O-2) to International Classification of Diseases and Related Health Problems, 10th Revision.(`http://www.cdc.gov/nchs/icd/icd10cm.htm`)

A.2. Análisis descriptivo

Variable	Categorías	Frecuencia (%)
Raza	Blanco	43272 (84.36 %)
	Negro	3802 (7.4 %)
	Indio americano, nativo de alaska	364 (0.7 %)
	Chino	744 (1.45 %)
	Japones	756 (1.47 %)
	Filipino	855 (1.66 %)
	Hawaiano	499 (0.97 %)
	Coreano	155 (0.3 %)
	Vietnamita	140 (0.27 %)
	Indio asiático	124 (0.24 %)
	Otros sitios de asía	162 (0.31 %)
	Otros	421 (0.69 %)
	Missing	94 (0.18 %)
Número de tumores primarios	0	42765 (83.22 %)
	1	1103 (2.14 %)
	2	6339 (12.34 %)
	3	983 (1.9 %)
	4	166 (0.32 %)
	5	22 (0.04 %)
	6	7 (0.01 %)
	8	1 (0.001 %)
	10	1 (0.001 %)
	Missing	1 (0.001 %)
Tratamiento	None	46408 (90.31 %)
	Beam radiation	3952 (7.69 %)
	Radioisotopes	5 (0.01 %)
	Combination of 1 with 2 or 3	15 (0.03 %)
	Radiation	33 (0.06 %)
	Other radiation	5 (0.01 %)
	Patient?s guardian refused radiation therapy	52 (0.1 %)
	Radiation recommended, unknown if administered	69 (0.13 %)
	Unknown	849 (1.65 %)

Tabla A.1: Tabla descriptiva de las variables raza, número de tumores primarios y tratamiento.

A.3. Función de riesgo

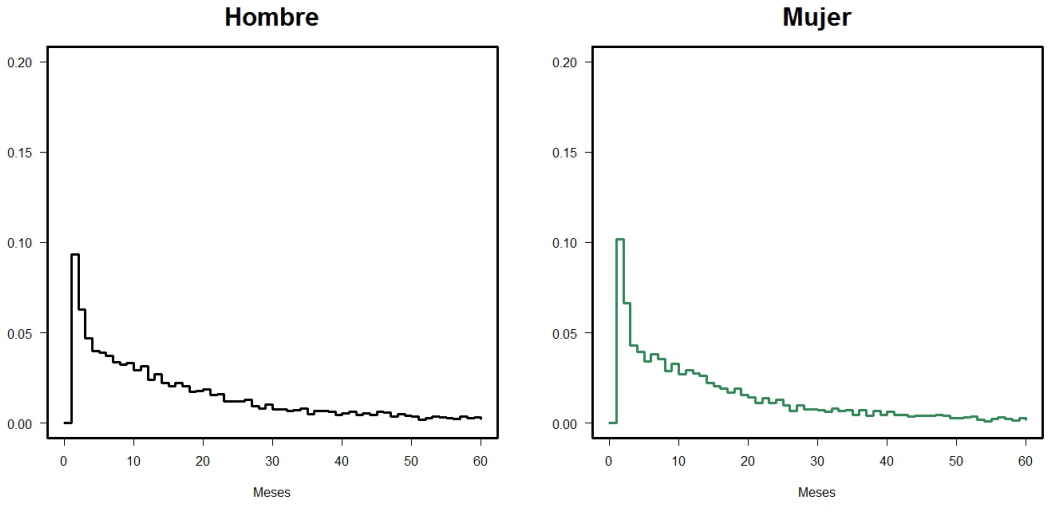

Figura A.1: Función de riesgo según el género. A la izquierda esta la curva para hombres y a la derecha mujeres.

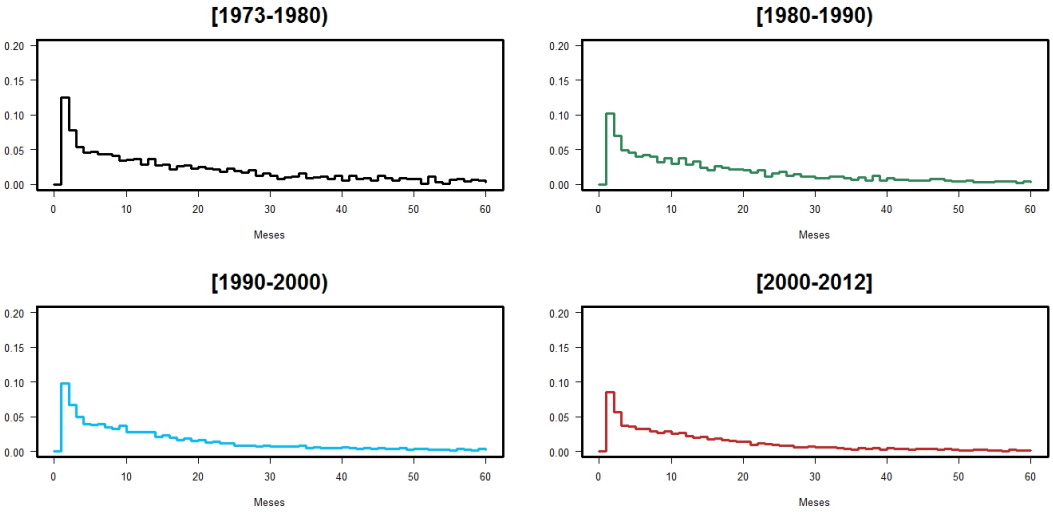

Figura A.2: Función de riesgo según el año del diagnóstico. De izquierda a derecha y de arriba abajo son, entre los años 1973 y 1980, entre 1980 y 1990, entre 1990 y 2000 y entre los años 2000 y 2012.

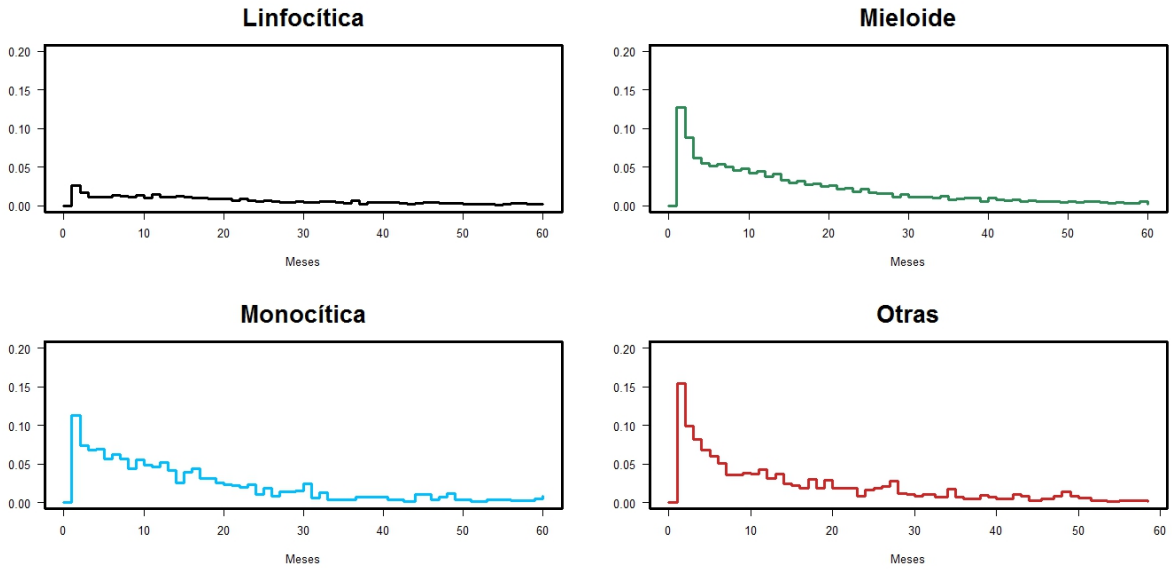

Figura A.3: Función de riesgo según e tipo de leucemia aguda. De izquierda a derecha y de arriba abajo son, linfocítica aguda, mieloide aguda, monocítica aguda y otro tipo de agudas.

Apéndice B

Código de R

```
library(descr)
library(survival)
library(rms)

##Leemos la base de datos que contiene las enfermedades linfoma, leucemia y mieloma.

linfleuc<-read.fwf("LYMYLEUK_1973.TXT",header = FALSE,fill = TRUE,
widths=c(8,10,1,2,1,1,1,3,4,2,2,4,4,1,4,1,4,1,1,1,1,3,2,2,1,2,2,13,2,4,1,1,1,1,3,3,3,2
,3,3,3,3,3,3,3,2,2,2,2,1,1,1,1,1,6,6,6,2,1,1,2,1,1,1,1,2,1,1,2,1,1,1,1,1,1,1,1,1,1,1
,1,1,2,5,4,4,3,3,1,2,2,3,1,1,1,1,2,2,1,1,2,1,5,5,5,1,1,1,2,2,1,1,1,1,1,1,1,1,2,3,3,3,3
,3,3,1,4,1,4,1,1,3,3,3,3,2,2,2,2,3,3,3,1,1,1))

#La variable V91 nos da la información sobre el tipo de enfermedad que tienen los individuos
#y la usaremos #para crear la variable grupo que separará las enfermedades de linfoma
#, leucemia y mieloma.

linfleuc$V91<-as.factor(linfleuc$V91)
for(i in 1:367088){
if (linfleuc$V91[i]==40|linfleuc$V91[i]==41|linfleuc$V91[i]==42|linfleuc$V91[i]==43
     |linfleuc$V91[i]==44){
       linfleuc$grupo[i]=1}
else if (linfleuc$V91[i]==49|linfleuc$V91[i]==50|linfleuc$V91[i]==51
        |linfleuc$V91[i]==52|linfleuc$V91[i]==53){
             linfleuc$grupo[i]=2 }
else if (linfleuc$V91[i]==45){
             linfleuc$grupo[i]=3}
else if (linfleuc$V91[i]==98){
             linfleuc$grupo[i]=4}}

linfleuc$grupo<-as.factor(linfleuc$grupo)
```

```
#Seleccionamos solo los casos de leucemia ya que es la enfermedad que vamos a analizar.
leuc<-subset(linfleuc,linfleuc$grupo==2)

#Tenemos todos los casos de leucemia que son 121280 y 147 variables.

#Creamos una variable  que se llamará 'tipo' para separar los tipos de leucemia, aguda(1)
#, crónica(2) y otras(3).

for (i in 1:121280){
if(leuc$V85[i]==35011|leuc$V85[i]==35021|leuc$V85[i]==35041|leuc$V85[i]==35031){
       leuc$tipo[i]=1}
else if (leuc$V85[i]==35012|leuc$V85[i]==35022){
       leuc$tipo[i]=2}
else if (leuc$V85[i]==35013|leuc$V85[i]==35023|leuc$V85[i]==35043){
       leuc$tipo[i]=3}
else {leuc$tipo[i]=999}}

leuc$tipo<-as.factor(leuc$tipo)

#Hacemos subset para quedarnos solo con la leucemia AGUDA, que es la que analizaremos.
aguda<-subset(leuc,tipo==1)

#Hacemos un summary para ver que variables tienen un porcentaje muy elevado de missing y
#las borramos.
summary(aguda)

aguda$V9<-NULL;aguda$V11<-NULL;aguda$V22<-NULL;aguda$V23<-NULL;aguda$V24<-NULL;
aguda$V25<-NULL;aguda$V30<-NULL;aguda$V35<-NULL;aguda$V36<-NULL;aguda$V37<-NULL;
aguda$V38<-NULL;aguda$V39<-NULL;aguda$V40<-NULL;aguda$V41<-NULL;aguda$V42<-NULL;
aguda$V43<-NULL;aguda$V44<-NULL;aguda$V45<-NULL;aguda$V46<-NULL;aguda$V47<-NULL;
aguda$V48<-NULL;aguda$V49<-NULL;aguda$V50<-NULL;aguda$V51<-NULL;aguda$V52<-NULL;
aguda$V53<-NULL;aguda$V54<-NULL;aguda$V55<-NULL;aguda$V56<-NULL;aguda$V57<-NULL;
aguda$V58<-NULL;aguda$V59<-NULL;aguda$V60<-NULL;aguda$V61<-NULL;aguda$V66<-NULL;
aguda$V67<-NULL;aguda$V68<-NULL;aguda$V70<-NULL;aguda$V71<-NULL;aguda$V72<-NULL;
aguda$V73<-NULL;aguda$V74<-NULL;aguda$V75<-NULL;aguda$V76<-NULL;aguda$V77<-NULL;
aguda$V78<-NULL;aguda$V79<-NULL;aguda$V80<-NULL;aguda$V81<-NULL;aguda$V82<-NULL;
aguda$V91<-NULL;aguda$V98<-NULL;aguda$V99<-NULL;aguda$V100<-NULL;aguda$V101<-NULL;
aguda$V109<-NULL;aguda$V114<-NULL;aguda$V115<-NULL;aguda$V116<-NULL;aguda$V121<-NULL;
aguda$V122<-NULL;aguda$V123<-NULL;aguda$V124<-NULL;aguda$V125<-NULL;aguda$V126<-NULL;
aguda$V127<-NULL;aguda$V132<-NULL;aguda$V133<-NULL;aguda$V134<-NULL;aguda$V135<-NULL;
aguda$V136<-NULL;aguda$V137<-NULL;aguda$V138<-NULL;aguda$V139<-NULL;aguda$V140<-NULL;
aguda$V141<-NULL;aguda$V142<-NULL;aguda$V143<-NULL;aguda$V143<-NULL;aguda$V144<-NULL;
aguda$V146<-NULL;aguda$grupo<-NULL;aguda$tipo<-NULL

#Hacemos factor las variables categóricas y les asignamos el código de valores perdidos.
```

```
aguda$V1<-as.factor(aguda$V1)
aguda$V2<-as.factor(aguda$V2)
aguda$V3<-as.factor(aguda$V3)
aguda$V3[aguda$V3 == 9] <- NA
aguda$V4<-as.factor(aguda$V4)
aguda$V4[aguda$V4 == 99] <- NA
aguda$V5<-as.factor(aguda$V5)
aguda$V6<-as.factor(aguda$V6)
aguda$V7<-as.factor(aguda$V7)
aguda$V8[aguda$V8==999]<- NA
aguda$V10<-as.factor(aguda$V10)
aguda$V10[ aguda$V10 == 99] <- NA
aguda$V10[ aguda$V10 == 88] <- NA
aguda$V13<-as.factor(aguda$V13)
aguda$V14<-as.factor(aguda$V14)
aguda$V15<-as.factor(aguda$V15)
aguda$V16<-as.factor(aguda$V16)
aguda$V17<-as.factor(aguda$V17)
aguda$V18<-as.factor(aguda$V18)
aguda$V19<-as.factor(aguda$V19)
aguda$V19[aguda$V19 == 9] <- NA
aguda$V20<-as.factor(aguda$V20)
aguda$V20[aguda$V20 == 9] <- NA
aguda$V21<-as.factor(aguda$V21)
aguda$V26<-as.factor(aguda$V26)
aguda$V26[aguda$V26 == 99] <- NA
aguda$V27<-as.factor(aguda$V27)
aguda$V27[aguda$V27 == 99] <- NA
aguda$V28<-as.factor(aguda$V28)
aguda$V29<-as.factor(aguda$V29)
aguda$V31<-as.factor(aguda$V31)
aguda$V32<-as.factor(aguda$V32)
aguda$V32[aguda$V32 == 9] <- NA
aguda$V33<-as.factor(aguda$V33)
aguda$V33[aguda$V33 == 9] <- NA
aguda$V34<-as.factor(aguda$V34)
aguda$V34[aguda$V34 == 9] <- NA
aguda$V62<-as.factor(aguda$V62)
aguda$V62[aguda$V62==8]<- NA
aguda$V62[aguda$V62==9]<- NA
aguda$V63<-as.factor(aguda$V63)
aguda$V63[aguda$V63==9]<- NA
aguda$V64<-as.factor(aguda$V64)
aguda$V64[aguda$V64 == 9] <- NA
```

```
aguda$V65<-as.factor(aguda$V65)
aguda$V65[aguda$V65 == 9] <- NA
aguda$V69<-as.factor(aguda$V69)
aguda$V83<-as.factor(aguda$V83)
aguda$V84<-as.factor(aguda$V84)
aguda$V84[aguda$V84==99]<- NA
aguda$V85<-as.factor(aguda$V85)
aguda$V85[aguda$V85==99999]<- NA
aguda$V86<-as.factor(aguda$V86)
aguda$V87<-as.factor(aguda$V87)
aguda$V88<-as.factor(aguda$V88)
aguda$V89<-as.factor(aguda$V89)
aguda$V90<-as.factor(aguda$V90)
aguda$V92<-as.factor(aguda$V92)
aguda$V92[aguda$V92 == 98] <- NA
aguda$V93<-as.factor(aguda$V93)
aguda$V94<-as.factor(aguda$V94)
aguda$V94[aguda$V94==9]<- NA
aguda$V95<-as.factor(aguda$V95)
aguda$V95[aguda$V95==9]<- NA
aguda$V96<-as.factor(aguda$V96)
aguda$V97<-as.factor(aguda$V97)
aguda$V103<-as.factor(aguda$V103)
aguda$V104<-as.factor(aguda$V104)
aguda$V105<-as.factor(aguda$V105)
aguda$V105[aguda$V105==99999]<- NA
aguda$V106<-as.factor(aguda$V106)
aguda$V107<-as.factor(aguda$V107)
aguda$V108<-as.factor(aguda$V108)
aguda$V110<-as.factor(aguda$V110)
aguda$V110[aguda$V110==99]<- NA
aguda$V111<-as.factor(aguda$V111)
aguda$V111[aguda$V111==99]<- NA
aguda$V112<-as.factor(aguda$V112)
aguda$V112[aguda$V112==9]<- NA
aguda$V113<-as.factor(aguda$V113)
aguda$V113[aguda$V113==9]<- NA
aguda$V117<-as.factor(aguda$V117)
aguda$V118<-as.factor(aguda$V118)
aguda$V118[aguda$V118==9]<- NA
aguda$V119<-as.factor(aguda$V119)
aguda$V119[aguda$V119==9]<- NA
aguda$V120<-as.factor(aguda$V120)
aguda$V129<-as.factor(aguda$V129)
aguda$V129[aguda$V129== 9] <- NA
```

```
aguda$V131<-as.factor(aguda$V131)
aguda$V131[aguda$V131==9] <- NA
aguda$V145<-as.factor(aguda$V145)
aguda$V145[aguda$V145==9] <- NA
aguda$V128[aguda$V128==9999]<- NA

#Volvemos a mirar los missing de las variables que nos quedan.
summary(aguda)
aguda$V15<-NULL;aguda$V19<-NULL;aguda$V26<-NULL;aguda$V27<-NULL;aguda$V28<-NULL;
aguda$V29<-NULL;aguda$V32<-NULL;aguda$V33<-NULL;aguda$V34<-NULL;aguda$V64<-NULL;
aguda$V92<-NULL;aguda$V108<-NULL;aguda$V111<-NULL;aguda$V118<-NULL;
aguda$V119<-NULL;aguda$V145<-NULL

#A continuación eliminamos las variables que tienen todos o casi todos los datos en la misma
#categoría, ya que esa información no aporta nada para el análisis.
summary(aguda)
aguda$V4<-NULL;aguda$V5<-NULL;aguda$V6<-NULL;aguda$V10<-NULL;aguda$V13<-NULL;aguda$V14<-NULL;
aguda$V16<-NULL;aguda$V18<-NULL;aguda$V20<-NULL;aguda$V21<-NULL;aguda$V62<-NULL;
aguda$V63<-NULL;aguda$V65<-NULL;aguda$V69<-NULL;aguda$V83<-NULL;aguda$V84<-NULL;
aguda$V86<-NULL;aguda$V87<-NULL;aguda$V88<-NULL;aguda$V89<-NULL;aguda$V90<-NULL;
aguda$V93<-NULL;aguda$V94<-NULL;aguda$V95<-NULL;aguda$V96<-NULL;aguda$V97<-NULL;
aguda$V102<-NULL;aguda$V103<-NULL;aguda$V106<-NULL;aguda$V110<-NULL;aguda$V117<-NULL;
aguda$V120<-NULL;aguda$V129<-NULL;aguda$V130<-NULL;aguda$V131<-NULL

#Eliminamos las filas donde las variable supervivencia es un valor perdido.
aguda<-aguda[complete.cases(aguda[,15]),]

#Renombramos las variables

names(aguda)[1:15] <- c( "id_num", "id", "DX","sex","ediag","adiag","hist_o3"
,"EOD","prim_recod","condado","causa","vital"
,"cemuerte","other","msurv")

#Categorizamos las variables adiag, ediag y tipo de leucemia.

##############
# TIPO DE LEUCEMIA AGUDA
##################

for(i in 1:50572){
if (aguda$prim_recod[i]==35011){
aguda$tipo_aguda[i]=1}
else if (aguda$prim_recod[i]==35021){
aguda$tipo_aguda[i]=2}
else if (aguda$prim_recod[i]==35031){
```

```
aguda$tipo_aguda[i]=3}
else if (aguda$prim_recod[i]==35041){
aguda$tipo_aguda[i]=4}
}
aguda$tipo_aguda<-as.factor(aguda$tipo_aguda)

# tipo_aguda==1 --> linfocítica
# tipo_aguda==2 --> mieloide
# tipo_aguda==3 --> monocítica
# tipo_aguda==4 --> otras

#Eliminamos la variable que no está categorizada.
aguda$prim_recod<-NULL

###############
# EDAD AL DIAGNÓSTICO
##################

aguda$ediag_categ[ aguda$ediag >= 0 & aguda$ediag < 14]<- "1"
aguda$ediag_categ[ aguda$ediag >=14 & aguda$ediag < 45]<- "2"
aguda$ediag_categ[ aguda$ediag >= 45 & aguda$ediag < 65]<- "3"
aguda$ediag_categ[ aguda$ediag >= 65]<- "4"

#Eliminamos la variable que no está categorizada.
aguda$ediag<-NULL

###############
# AÑO DEL DIAGNÓSTICO
##################

aguda$adiag_categ[ aguda$adiag >= 1973 & aguda$adiag < 1980]<- "1"
aguda$adiag_categ[ aguda$adiag >=1980 & aguda$adiag < 1990]<- "2"
aguda$adiag_categ[ aguda$adiag >= 1990 & aguda$adiag < 2000]<- "3"
aguda$adiag_categ[ aguda$adiag >= 2000]<- "4"

#Eliminamos la variable que no está categorizada.
aguda$adiag<-NULL

#Hacemos la variable para los pacientes curados.
#1-->están enfermos o no han pasado aun el periodo para considerarlos curados
#2-->están CURADOS

for(i in 1:50572){
if(aguda$msurv[i]>60){
     aguda$curados[i]<-2}
```

```
else  {aguda$curados[i]<-1} }

aguda$curados<-as.factor(aguda$curados)

###CENSURA
#1--> muerto por leucemia aguda o causa atribuible en los primeros 5 años desde el diagnóstico
#0-->vivo o muerto por otra causa que no sea leucemia

for(i in 1:50572){
if((aguda$curados[i]!=2 & aguda$causa[i]==35011)|(aguda$curados[i]!=2 & aguda$causa[i]==35021)
|(aguda$curados[i]!=2 & aguda$causa[i]==35031)|(aguda$curados[i]!=2 & aguda$causa[i]==35041)
|(aguda$curados[i]!=2 & aguda$causa[i]==5000)|(aguda$curados[i]!=2 & aguda$causa[i]==50010)
|(aguda$curados[i]!=2 & aguda$causa[i]==50030)|(aguda$curados[i]!=2 & aguda$causa[i]==50040)
|(aguda$curados[i]!=2 & aguda$causa[i]==50120)|(aguda$curados[i]!=2 & aguda$causa[i]==50140)){
aguda$censura[i]<-1 }
else {aguda$censura[i]<-0}}

#Par trabajar la supervicvenca le sumaremos 0.5 y así supondremos que a lo largo del mes
# han muerto a medados ya que no está especificado dentro del mes en que momento pasó

aguda$msurv<-aguda$msurv+0.5

#############################################################################
######      KAPLAN Y MEIER
#############################################################################

sagu <- with(aguda, Surv(msurv, censura))
surv <- survfit(sagu~1)

windows()
par(mar=c(5,5,4,2),las=1)
plot(surv , lwd=3, xlab = 'Meses', ylab =expression(bold(hat(S)(t))),mark.time=F,
       conf.int=F,xmax=60)
abline(h=0.5,col="Red",lwd=3)
title('Tiempo de supervivencia')

##Curva a lo dos años (24 meses)
windows()
par(mar=c(5,5,4,2),las=1)
plot(surv , lwd=3, xlab = 'Meses', ylab =expression(bold(hat(S)(t)))
     ,mark.time=F,conf.int=F,xmax=24)
abline(h=0.5,col="Red",lwd=3)
title('Tiempo de supervivencia')

#############
```

```
##Género
#################

surv1<-survfit(sagu~sex,aguda)

windows()
par(mar=c(5,5,4,2))
plot(surv1, col=3:4, xlab='Tiempo hasta la muerte [meses]', ylab=expression(bold(hat(S)(t)))
      , mark.time=F, lty=1:2, lwd=3,xmax=60)
title('Supervivencia según género')
legend('topright', c("Hombre","Mujer"),bty='n', col=3:4, lty=1:2, lwd=3)

######################
### EDAD
########################

surv2<-survfit(sagu~ediag_categ,aguda)

windows()
par(mar=c(5,5,4,2))
plot(surv2, main = "",col=3:6, xlab='Tiempo hasta la muerte[meses]',
        ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Supervivencia según la edad al diagnóstico', cex = 0.7))
legend('topright', c("< 14","[14,45)","[45,65)",">= 65"),bty='n',cex = 0.55, col=3:6
       , lty=1:4, lwd=3)

########################
### TIPO DE LEUCEMIA AGUDA
########################

surv3<-survfit(sagu~tipo_aguda,aguda)

windows()
par(mar=c(5,5,4,2))
plot(surv3,main = "", col=3:6, xlab='Tiempo hasta la muerte [meses]',
          ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Supervivencia según el tipo de leucemia',cex = 0.8))
legend('topright', c("lymphocytic","myeloid","monocytic","otras"),bty='n',cex = 0.7, col=3:6
            , lty=1:4, lwd=3)

########################
### AÑO DEL DIAGNÓSTICO
########################

surv4<-survfit(sagu~adiag_categ,aguda)
```

```
windows()
par(mar=c(5,5,4,2))
plot(surv4, col=3:6, xlab='Tiempo hasta la muerte [meses]', ylab=expression(bold(hat(S)(t)))
        , mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Supervivencia según año del diagnóstico',cex = 0.8))
legend('topright', c("1973-1980","1980-1990","1990-2000","2000-2012"),bty='n',cex = 0.7
        , col=3:6, lty=1:4, lwd=3)

########################
### SEGÚN CADA CATEGORÍA DE EDAD
########################

menos14<-subset(aguda,ediag_categ==1)
entre14y45<-subset(aguda,ediag_categ==2)
entre45y65<-subset(aguda,ediag_categ==3)
mayor65<-subset(aguda,ediag_categ==4)

sagu14<-with(menos14, Surv(msurv, censura))
sagu14y45<-with(entre14y45, Surv(msurv, censura))
sagu45y65<-with(entre45y65, Surv(msurv, censura))
sagu65<-with(mayor65, Surv(msurv, censura))

#Tipo de aguda

svf14<-survfit(sagu14~tipo_aguda,menos14)
svf14y45<-survfit(sagu14y45~tipo_aguda,entre14y45)
svf45y65<-survfit(sagu45y65~tipo_aguda,entre45y65)
svf65<-survfit(sagu65~tipo_aguda,mayor65)

windows()
par(mfrow=c(2,2))
plot(svf14,main = "", col=3:6, xlab='Tiempo hasta la muerte [meses]'
       , ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Menores de 14 años',cex = 0.8))
legend('topright', c("lymphocytic","myeloid","monocytic","otras"),bty='n',cex = 0.7, col=3:6
        , lty=1:4, lwd=3)

plot(svf14y45,main = "", col=3:6, xlab='Tiempo hasta la muerte [meses]'
       , ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Edad entre 14 y 45 años',cex = 0.8))
legend('topright', c("lymphocytic","myeloid","monocytic","otras"),bty='n',cex = 0.7, col=3:6
         , lty=1:4, lwd=3)

plot(svf45y65,main = "", col=3:6, xlab='Tiempo hasta la muerte [meses]'
```

```
         , ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Edad entre 45 y 65 años',cex = 0.8))
legend('topright', c("lymphocytic","myeloid","monocytic","otras"),bty='n',cex = 0.7, col=3:6
           , lty=1:4, lwd=3)

plot(svf65,main = "", col=3:6, xlab='Tiempo hasta la muerte [meses]'
         , ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Mayores de 65 años',cex = 0.8))
legend('topright', c("lymphocytic","myeloid","monocytic","otras"),bty='n',cex = 0.7, col=3:6
         , lty=1:4, lwd=3)

#########################
### SEGÚN TIPO DE LEUCEMIA
#########################

lympho<-subset(aguda,tipo_aguda==1)
myelo<-subset(aguda,tipo_aguda==2)
monocy<-subset(aguda,tipo_aguda==3)
otras<-subset(aguda,tipo_aguda==4)

sagulympho<-with(lympho, Surv(msurv, censura))
sagumyelo<-with(myelo, Surv(msurv, censura))
sagumonocy<-with(monocy, Surv(msurv, censura))
saguotras<-with(otras, Surv(msurv, censura))

#EDAD
svflympho<-survfit(sagulympho~ediag_categ,lympho)
svfmyelo<-survfit(sagumyelo~ediag_categ,myelo)
svfmonocy<-survfit(sagumonocy~ediag_categ,monocy)
svfotras<-survfit(saguotras~ediag_categ,otras)

windows()
par(mfrow=c(2,2))
plot(svflympho, main = "",col=3:6, xlab='Tiempo hasta la muerte[meses]'
    , ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Lymphocytica', cex = 0.7))
legend('topright', c("< 14","[14,45)","[45,65)",">= 65"),bty='n',cex = 0.55, col=3:6
         , lty=1:4, lwd=3)

plot(svfmyelo, main = "",col=3:6, xlab='Tiempo hasta la muerte[meses]'
     , ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Myeloid', cex = 0.7))
legend('topright', c("< 14","[14,45)","[45,65)",">= 65"),bty='n',cex = 0.55, col=3:6
         , lty=1:4, lwd=3)
```

```
plot(svfmonocy, main = "",col=3:6, xlab='Tiempo hasta la muerte[meses]'
      , ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Monocytic', cex = 0.7))
legend('topright', c("< 14","[14,45)","[45,65)",">= 65"),bty='n',cex = 0.55, col=3:6
          , lty=1:4, lwd=3)

plot(svfotras, main = "",col=3:6, xlab='Tiempo hasta la muerte[meses]'
      , ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Otras leucemias agudas', cex = 0.7))
legend('topright', c("< 14","[14,45)","[45,65)",">= 65"),bty='n',cex = 0.55, col=3:6
      , lty=1:4, lwd=3)


#Año DIAGNÓSTICO
svflympho2<-survfit(sagulympho~adiag_categ,lympho)
svfmyelo2<-survfit(sagumyelo~adiag_categ,myelo)
svfmonocy2<-survfit(sagumonocy~adiag_categ,monocy)
svfotras2<-survfit(saguotras~adiag_categ,otras)

windows()
par(mfrow=c(2,2))
plot(svflympho2, col=3:6, xlab='Tiempo hasta la muerte [meses]'
     , ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Lymphocytica',cex = 0.8))
legend('topright', c("1973-1980","1980-1990","1990-2000","2000-2012"),bty='n',cex = 0.7
        , col=3:6, lty=1:4, lwd=3)

plot(svfmyelo2, col=3:6, xlab='Tiempo hasta la muerte [meses]'
      , ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Myeloid',cex = 0.8))
legend('topright', c("1973-1980","1980-1990","1990-2000","2000-2012"),bty='n',cex = 0.7
        , col=3:6, lty=1:4, lwd=3)

plot(svfmonocy2, col=3:6, xlab='Tiempo hasta la muerte [meses]'
       , ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Monocytic',cex = 0.8))
legend('topright', c("1973-1980","1980-1990","1990-2000","2000-2012"),bty='n',cex = 0.7
        , col=3:6, lty=1:4, lwd=3)

plot(svfotras2, col=3:6, xlab='Tiempo hasta la muerte [meses]'
      , ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Otras leucemias agudas',cex = 0.8))
legend('topright', c("1973-1980","1980-1990","1990-2000","2000-2012"),bty='n',cex = 0.7
        , col=3:6, lty=1:4, lwd=3)
```

```
#########################
### SEGÚN AÑO DEL DIAGNÓSTICO
#########################

entre70y80<-subset(aguda,adiag_categ==1)
entre80y90<-subset(aguda,adiag_categ==2)
entre90y2000<-subset(aguda,adiag_categ==3)
mas2000<-subset(aguda,adiag_categ==4)

saguentre70y80<-with(entre70y80, Surv(msurv, censura))
saguentre80y90<-with(entre80y90, Surv(msurv, censura))
saguentre90y2000<-with(entre90y2000, Surv(msurv, censura))
sagumas2000<-with(mas2000, Surv(msurv, censura))

#TIPO
svf70y80<-survfit(saguentre70y80~tipo_aguda,entre70y80)
svf80y90<-survfit(saguentre80y90~tipo_aguda,entre80y90)
svf90y2000<-survfit(saguentre90y2000~tipo_aguda,entre90y2000)
svf2000<-survfit(sagumas2000~tipo_aguda,mas2000)

windows()
par(mfrow=c(2,2))
plot(svf70y80,main = "", col=3:6, xlab='Tiempo hasta la muerte [meses]',
     ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Entre el año 1970 7 1980',cex = 0.8))
legend('topright', c("lymphocytic","myeloid","monocytic","otras"),bty='n',cex = 0.7, col=3:6,
           lty=1:4, lwd=3)

plot(svf80y90,main = "", col=3:6, xlab='Tiempo hasta la muerte [meses]',
         ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Entre el año 1980 y 1990',cex = 0.8))
legend('topright', c("lymphocytic","myeloid","monocytic","otras"),bty='n',cex = 0.7, col=3:6,
             lty=1:4, lwd=3)

plot(svf90y2000,main = "", col=3:6, xlab='Tiempo hasta la muerte [meses]',
          ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Entre el año 1990 y 2000',cex = 0.8))
legend('topright', c("lymphocytic","myeloid","monocytic","otras"),bty='n',cex = 0.7, col=3:6,
               lty=1:4, lwd=3)

plot(svf2000,main = "", col=3:6, xlab='Tiempo hasta la muerte [meses]',
          ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('A partir del año 2000',cex = 0.8))
legend('topright', c("lymphocytic","myeloid","monocytic","otras"),bty='n',cex = 0.7, col=3:6,
             lty=1:4, lwd=3)
```

```
#EDAD
svf70y80_2<-survfit(saguentre70y80~ediag_categ,entre70y80)
svf80y90_2<-survfit(saguentre80y90~ediag_categ,entre80y90)
svf90y2000_2<-survfit(saguentre90y2000~ediag_categ,entre90y2000)
svf2000_2<-survfit(sagumas2000~ediag_categ,mas2000)

windows()
par(mfrow=c(2,2))
plot(svf70y80_2, main = "",col=3:6, xlab='Tiempo hasta la muerte[meses]',
       ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Entre el año 1970 7 1980', cex = 0.7))
legend('topright', c("< 14","[14,45)","[45,65)",">= 65"),bty='n',cex = 0.55, col=3:6
     , lty=1:4,  lwd=3)

plot(svf80y90_2, main = "",col=3:6, xlab='Tiempo hasta la muerte[meses]',
            ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Entre el año 1980 y 1990', cex = 0.7))
legend('topright', c("< 14","[14,45)","[45,65)",">= 65"),bty='n',cex = 0.55, col=3:6
             , lty=1:4, lwd=3)

plot(svf90y2000_2, main = "",col=3:6, xlab='Tiempo hasta la muerte[meses]',
           ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('Entre el año 1990 y 2000', cex = 0.7))
legend('topright', c("< 14","[14,45)","[45,65)",">= 65"),bty='n',cex = 0.55, col=3:6
          , lty=1:4, lwd=3)

plot(svf2000_2, main = "",col=3:6, xlab='Tiempo hasta la muerte[meses]',
          ylab=expression(bold(hat(S)(t))), mark.time=F, lty=1:4, lwd=3,xmax=60)
title(main = list('A partir del año 2000', cex = 0.7))
legend('topright', c("< 14","[14,45)","[45,65)",">= 65"),bty='n',cex = 0.55, col=3:6
         , lty=1:4, lwd=3)

##################################################
##  HAZARD
##################################################

library(muhaz)

#Actualizar la función kphaz.plot

kphaz.plot2<-function (fit, ...) {
if (any(names(fit) == "time") & (any(names(fit) == "haz"))) {
time <- fit$time
haz <- fit$haz}
```

```
else stop("Argument \"fit\" must be the result of a call to \"kphaz.fit\"")
if (length(time) != length(haz))
stop("Argument \"fit\" must be the result of a call to \"kphaz.fit\"")
qstrata <- any(names(fit) == "strata")
if (qstrata)
strata <- fit$strata
else strata <- rep(1, length(time))
if (length(strata) != length(haz))
stop("Argument \"fit\" must be the result of a call to \"kphaz.fit\"")
ustrata <- unique(strata)
good <- 1:length(ustrata)
for (i in 1:length(ustrata)) {
cur.strata <- ustrata[i]
ind <- strata == ustrata[i]
if (all(is.na(haz[ind] | is.nan(haz[ind]))))
good <- good[good != i]}
ustrata <- ustrata[good]
if (length(ustrata) < 1)
stop("No plots")
ind <- (strata == ustrata[1]) & (!is.nan(haz))
xmax <- max(time)
#ymax <- max(haz[((!is.nan(haz)) & (!is.na(haz)))])
ymax<-0.2
x <- time[ind]
y <- haz[ind]
if (min(x) > 0) {
x <- c(0, x)
y <- c(0, y)}
plot(x, y, xlim = c(0, xmax), ylim = c(0, ymax), xlab = "Meses",
ylab = "", type = "s", ...)
if (length(ustrata) > 1) {
for (i in 2:length(ustrata)) {
ind <- strata == ustrata[i]
x <- time[ind]
y <- haz[ind]
if (min(x) > 0) {
x <- c(0, x)
y <- c(0, y)}
lines(stepfun(x, y), lty = i)}}
invisible()}

kphaz.plot3<-function (fit, ...){
if (any(names(fit) == "time") & (any(names(fit) == "haz"))) {
time <- fit$time
haz <- fit$haz}
```

```
else stop("Argument \"fit\" must be the result of a call to \"kphaz.fit\"")
if (length(time) != length(haz))
stop("Argument \"fit\" must be the result of a call to \"kphaz.fit\"")
qstrata <- any(names(fit) == "strata")
if (qstrata)
strata <- fit$strata
else strata <- rep(1, length(time))
if (length(strata) != length(haz))
stop("Argument \"fit\" must be the result of a call to \"kphaz.fit\"")
ustrata <- unique(strata)
good <- 1:length(ustrata)
for (i in 1:length(ustrata)) {
cur.strata <- ustrata[i]
ind <- strata == ustrata[i]
if (all(is.na(haz[ind] | is.nan(haz[ind]))))
good <- good[good != i]}
ustrata <- ustrata[good]
if (length(ustrata) < 1)
stop("No plots")
ind <- (strata == ustrata[1]) & (!is.nan(haz))
xmax <- max(time)
ymax <- max(haz[((!is.nan(haz)) & (!is.na(haz)))])
x <- time[ind]
y <- haz[ind]
if (min(x) > 0) {
x <- c(0, x)
y <- c(0, y)}
plot(x, y, xlim = c(0, xmax), ylim = c(0, ymax), xlab = "Meses",
ylab = "", type = "s", ...)
if (length(ustrata) > 1) {
for (i in 2:length(ustrata)) {
ind <- strata == ustrata[i]
x <- time[ind]
y <- haz[ind]
if (min(x) > 0) {
x <- c(0, x)
y <- c(0, y)}
lines(stepfun(x, y), lty = i)}}
invisible()}

windows()
riesgo<-kphaz.fit(aguda$msurv,aguda$censura,q=1,method="nelson")
kphaz.plot3(riesgo,col="black", lwd=3,las=1)

#Segun tipo de aguda
```

```
fitlympho<-kphaz.fit(lympho$msurv,lympho$censura,q=1,method="nelson")
fitmyelo<-kphaz.fit(myelo$msurv,myelo$censura,q=1,method="nelson")
fitmonocy<-kphaz.fit(monocy$msurv,monocy$censura,q=1,method="nelson")
fitotras<-kphaz.fit(otras$msurv,otras$censura,q=1,method="nelson")

windows()
par(mfrow = c(2, 2), lwd = 3, las = 1,oma=c(0,0, 1,0),cex.main=2)
kphaz.plot2(fitlympho,col="black", lwd=3)
title("Linfocítica")
kphaz.plot2(fitmyelo,col="seagreen", lwd=3)
title("Mieloide")
kphaz.plot2(fitmonocy, col="deepskyblue",lwd=3)
title("Monocítica")
kphaz.plot2(fitotras, col="firebrick3", lwd=3)
title("Otras")

#Según edad

fitmenos14<-kphaz.fit(menos14$msurv,menos14$censura,q=1,method="nelson")
fitentre14y45<-kphaz.fit(entre14y45$msurv,entre14y45$censura,q=1,method="nelson")
fitentre45y65<-kphaz.fit(entre45y65$msurv,entre45y65$censura,q=1,method="nelson")
fitmayor65<-kphaz.fit(mayor65$msurv,mayor65$censura,q=1,method="nelson")

windows()
par(mfrow = c(2, 2), lwd = 3, las = 1,oma=c(0,0, 1,0),cex.main=2)
kphaz.plot2(fitmenos14,col="black", lwd=3)
title("Menores de 14 años")
kphaz.plot2(fitentre14y45,col="seagreen", lwd=3)
title("Edad entre 14-45 años")
kphaz.plot2(fitentre45y65, col="deepskyblue",lwd=3)
title("Edad entre 45 y 65 años")
kphaz.plot2(fitmayor65, col="firebrick3", lwd=3)
title("Mayores de 65 años")

#Segun años

fitentre70y80<-kphaz.fit(entre70y80$msurv,entre70y80$censura,q=1,method="nelson")
fitentre80y90<-kphaz.fit(entre80y90$msurv,entre80y90$censura,q=1,method="nelson")
fitentre90y2000<-kphaz.fit(entre90y2000$msurv,entre90y2000$censura,q=1,method="nelson")
fitmas2000<-kphaz.fit(mas2000$msurv,mas2000$censura,q=1,method="nelson")

windows()
par(mfrow = c(2, 2), lwd = 3, las = 1,oma=c(0,0, 1,0),cex.main=2)
kphaz.plot2(fitentre70y80,col="black", lwd=3)
title("[1973-1980)")
```

```
kphaz.plot2(fitentre80y90,col="seagreen", lwd=3)
title("[1980-1990)")
kphaz.plot2(fitentre90y2000, col="deepskyblue",lwd=3)
title("[1990-2000)")
kphaz.plot2(fitmas2000, col="firebrick3", lwd=3)
title("[2000-2012]")

#Según género

hombre<-subset(aguda,sex==1)
mujer<-subset(aguda,sex==2)
saguhombre<-with(hombre, Surv(msurv, censura))
sagujumer<-with(mujer, Surv(msurv, censura))
fithombre<-kphaz.fit(hombre$msurv,hombre$censura,q=1,method="nelson")
fitmujer<-kphaz.fit(mujer$msurv,mujer$censura,q=1,method="nelson")

windows()
par(mfrow = c(1, 2), lwd = 3, las = 1,oma=c(0,0, 1,0),cex.main=2)
kphaz.plot2(fithombre,col="black", lwd=3)
title("Hombre")
kphaz.plot2(fitmujer,col="seagreen", lwd=3)
title("Mujer")

#####################################################################
##### COX
#####################################################################
library(splines)

myesurv <- with(aguda, Surv(msurv, censura))

#Hacemos el modelo con las variables sexo, tipo de leucemia (dentro de las agudas),
# edad al diagnóstico y año del diagnóstico

cox1<-coxph(myesurv~sex+tipo_aguda+ediag_categ+adiag_categ,aguda)
summary(cox1)

ph.test<-cox.zph(cox1)

windows()
par(mfrow=c(2,3),las=1)
plot(ph.test,var=1,ylim=c(-2,2))
abline(h=0,col="red",lwd=2)
title("Mujer")
plot(ph.test,var=2,ylim=c(-2,2))
abline(h=0,col="red",lwd=2)
```

```
title("Leucemia mieloide aguda")
plot(ph.test,var=3,ylim=c(-2,2))
abline(h=0,col="red",lwd=2)
title("Leucemia monocítica aguda")
plot(ph.test,var=4,ylim=c(-2,2))
abline(h=0,col="red",lwd=2)
title("Otras leucemias agudas")
plot(ph.test,var=5,ylim=c(-5,5))
abline(h=0,col="red",lwd=2)
title("Edad entre 14-45 años")
plot(ph.test,var=6,ylim=c(-5,5))
abline(h=0,col="red",lwd=2)
title("Edad entre 45 y 65 años")

windows()
par(mfrow=c(2,3),las=1)
plot(ph.test,var=7,ylim=c(-5,5))
abline(h=0,col="red",lwd=2)
title("Mayores de 65 años")
plot(ph.test,var=8,ylim=c(-5,5))
abline(h=0,col="red",lwd=2)
title("Diagnóstico entre 1980 y 1990")
plot(ph.test,var=9,ylim=c(-5,5))
abline(h=0,col="red",lwd=2)
title("Diagnóstico entre 1990 y 2000")
plot(ph.test,var=10,ylim=c(-5,5))
abline(h=0,col="red",lwd=2)
title("Diagnóstico entre 2000 y 2012")

##############################
# COX estratificado por tipo de leucemia
##############################

coxEstr<-coxph(myesurv~sex+ediag_categ+adiag_categ+strata(tipo_aguda),aguda)
summary(coxEstr)
ph.test_Estr<-cox.zph(coxEstr)

windows()
par(mfrow=c(2,2),las=1)
plot(ph.test_Estr,var=1,ylim=c(-2,2))
abline(h=0,col="red",lwd=2)
title("Mujer")
plot(ph.test_Estr,var=2,ylim=c(-2,2))
abline(h=0,col="red",lwd=2)
title("Edad entre 14-45 años")
```

```
plot(ph.test_Estr,var=3,ylim=c(-2,2))
abline(h=0,col="red",lwd=2)
title("Edad entre 45 y 65 años")
plot(ph.test_Estr,var=4,ylim=c(-5,5))
abline(h=0,col="red",lwd=2)
title("Mayores de 65 años")

windows()
par(mfrow=c(2,2),las=1)
plot(ph.test_Estr,var=5,ylim=c(-5,5))
abline(h=0,col="red",lwd=2)
title("Diagnóstico entre 1980 y 1990")
plot(ph.test_Estr,var=6,ylim=c(-5,5))
abline(h=0,col="red",lwd=2)
title("Diagnóstico entre 1990 y 2000")
plot(ph.test_Estr,var=7,ylim=c(-5,5))
abline(h=0,col="red",lwd=2)
title("Diagnóstico entre 2000 y 2012")

#########################################################
######     MODELO ADITIVO DE AALEN
#########################################################

#install.packages("timereg")
library(timereg)

#Empezamos por un modelo básico y vamos probando hasta conseguir el óptimo

aalen1<-aalen(Surv(msurv, censura)~sex,aguda,max.time=60)
summary(aalen1)
par(mfrow=c(1,2))
plot(aalen1,hw.ci=2)

aalen2<-aalen(Surv(msurv, censura)~sex+tipo_aguda,aguda,max.time=60)
summary(aalen2)
par(mfrow=c(2,3))
plot(aalen2,hw.ci=2)

aalen3<-aalen(Surv(msurv, censura)~sex+tipo_aguda+ediag_categ,aguda,max.time=60)
par(mfrow=c(3,3))
plot(aalen3,hw.ci=2,ylim=c(-2,2))
summary(aalen3)

aalen4<-aalen(Surv(msurv, censura)~tipo_aguda+ediag_categ+adiag_categ,aguda,max.time=60)
par(mfrow=c(2,3))
```

```
plot(aalen4,hw.ci=2)
summary(aalen4)

aalen5<-aalen(Surv(msurv, censura)~sex+tipo_aguda+ediag_categ+adiag_categ,aguda,max.time=60)
summary(aalen5)

windows()
par(mfrow=c(2,3),las=1)
plot(aalen5,specific.comps=1)
plot(aalen5,specific.comps=2,mains=F,xlab="Tiempo (meses)",ylab="Coeficiente acumulativo")
title(main="Mujer")
plot(aalen5,specific.comps=3,mains=F,xlab="Tiempo (meses)",ylab="Coeficiente acumulativo")
title(main="Mieloide")
plot(aalen5,specific.comps=4,mains=F,xlab="Tiempo (meses)",ylab="Coeficiente acumulativo")
title(main="Monocítica")
plot(aalen5,specific.comps=5,mains=F,xlab="Tiempo (meses)",ylab="Coeficiente acumulativo")
title(main="Otras")

windows()
par(mfrow=c(2,3),las=1)
plot(aalen5,specific.comps=6,mains=F,xlab="Tiempo (meses)",ylab="Coeficiente acumulativo")
title(main="Edad entre 14-45 años")
plot(aalen5,specific.comps=7,mains=F,xlab="Tiempo (meses)",ylab="Coeficiente acumulativo")
title(main="Edad entre 45 y 65 años")
plot(aalen5,specific.comps=8,mains=F,xlab="Tiempo (meses)",ylab="Coeficiente acumulativo")
title(main="Mayores de 65 años")
plot(aalen5,specific.comps=9,mains=F,xlab="Tiempo (meses)",ylab="Coeficiente acumulativo")
title(main="Diagnóstico entre 1980 y 1990")
plot(aalen5,specific.comps=10,mains=F,xlab="Tiempo (meses)",ylab="Coeficiente acumulativo")
title(main="Diagnóstico entre 1990 y 2000")
plot(aalen5,specific.comps=11,mains=F,xlab="Tiempo (meses)",ylab="Coeficiente acumulativo")
title(main="Diagnóstico entre 2000 y 2012")

#Hacemos la parte semi-parametrica según los gráficos

aalen5.semi<-aalen(Surv(msurv, censura)~const(sex)+tipo_aguda+ediag_categ+const(adiag_categ)
,aguda,max.time=60)
summary(aalen5.semi)

windows()
par(mfrow=c(2,2),las=1)
plot(aalen5.semi,specific.comps=1,xlab="Tiempo (meses)",ylab="Coeficiente acumulativo")
plot(aalen5.semi,specific.comps=2,mains=F,xlab="Tiempo(meses)",ylab="Coeficiente acumulativo")
title(main="Mieloide")
plot(aalen5.semi,specific.comps=3,mains=F,xlab="Tiempo(meses)",ylab="Coeficiente acumulativo")
```

```
title(main="Monocítica")
plot(aalen5.semi,specific.comps=4,mains=F,xlab="Tiempo(meses)",ylab="Coeficiente acumulativo")
title(main="Otras")

windows()
par(mfrow=c(2,2),las=1)
plot(aalen5.semi,specific.comps=5,mains=F,xlab="Tiempo(meses)",ylab="Coeficiente acumulativo")
title(main="Edad entre 14-45 años")
plot(aalen5.semi,specific.comps=6,mains=F,xlab="Tiempo(meses)",ylab="Coeficiente acumulativo")
title(main="Edad entre 45 y 65 años")
plot(aalen5.semi,specific.comps=7,mains=F,xlab="Tiempo(meses)",ylab="Coeficiente acumulativo")
title(main="Mayores de 65 años")
```