TRABAJO DE FIN DE GRADO

Grado en Estadística

Título: Estudio de factores asociados a la supervivencia de pacientes con cáncer colorectal en Estados Unidos

Autor: Víctor Navarro Garcés

Director: Klaus Langohr

Departamento: Departamento de Estadística e Investigación Operativa

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Grado en Estadística
Universitat de Barcelona - Universitat politècnica de Catalunya

Estudio de factores asociados a la supervivencia de pacientes con cáncer colorrectal en Estados Unidos

Autor: Víctor Navarro Garcés
Director: Klaus Langohr

Junio, 2020

Barcelona

Resumen

El cáncer es una enfermedad que lleva muchos años siendo una de las principales causas de mortalidad del mundo. Se han obtenido una gran cantidad de datos del cáncer colorrectal en Estados Unidos para un período temporal que va desde 1975 hasta 2016. El objetivo de este trabajo ha sido estudiar los factores asociados a la supervivencia del cáncer colorrectal a través del Modelo de Cox y Modelo de Vida Acelerada. Al dividirse la base de datos en dos ejes temporales, una de pacientes diagnosticados entre los años 1975 y 1984 y otra con los diagnosticados entre 2004 y 2011, se ha comprobado como la supervivencia de este tipo de tumor ha mejorado para los tiempos más recientes. Además, se ha podido comprobar como las personas afroamericanas tienen una menor supervivencia. Como cabía esperar el riesgo de muerte en un plazo de cinco años desde el diagnóstico del tumor es mayor para las personas que carecen de seguro médico privado, las que tienen más edad, las que tienen un tamaño de tumor más grande y las que tienen un cáncer de grado más alto.

Palabras clave: Modelo de Cox, Modelo de Vida Acelerada, Cáncer Colorrectal, Análisis de Supervivencia.

Clasificación AMS: 62-07 (Análisis de datos), 62N01 (Modelos de datos censurados), 62N02 (Estimaciones), 62P10 (Aplicaciones a la biología y las ciencias médicas).

Abstract

Cancer is a disease that has been one of the world's leading causes of mortality for many years. A large amount of colorrectal cancer data has been obtained in the United States from 1975 to 2016. The objective of this work will be to study the factors associated with the survival of colorrectal cancer using the Cox Model and the Accelerated failure time model. Due to dividing the database into two smaller ones, one with patients diagnosed between 1975 and 1984 and the other with those diagnosed between 2004 and 2011, it has been found that the survival of this type of tumour has improved. Furthermore, it has been verified that African-American people have lower survival. As expected, the 5-year risk of death is higher for people without private health insurance, those who are older, those with larger tumour size, and those with higher-grade cancer.

Key words: Cox Model, Accelerated failure time model, colorrectal Cancer, Survival Analysis.

AMS classification: 62-07 (Data analysis), 62N01 (Censored data models), 62N02 (Estimations), 62P10 (Applications to biology and medical sciences).

Índice General

Índice de Tablas

Índice de códigos

Índice de Gráficos

Capítulo I

Introducción

1.1. Motivación

En los últimos años, el cáncer ha sido una de las principales causas de mortalidad en todo el mundo. A pesar de la disminución de las tasas de mortalidad gracias al avance tecnológico y científico, esta enfermedad sigue siendo una de las más mortíferas del planeta.

Según datos estimados por el proyecto *"GLOBOCAN 2018"* (ver *Las Cifras del Cáncer en España 2020* [14]) ha habido aproximadamente 18,1 millones de casos nuevos en el mundo. Observando las estimaciones poblacionales se prevé un aumento en las dos próximas décadas, incluso llegando a 29,5 millones en el año 2040.

En el caso de España, el cáncer también es una de las principales causas de mortalidad. Según los cálculos de la *Red Española de Registros de Cáncer* se prevé que el 2020 acabe con un número parecido de casos nuevos de cáncer que en 2019 (aproximadamente 277000 casos).

A nivel mundial, según datos estimados por el proyecto *"GLOBOCAN 2018 "*, el cáncer colorrectal (10,2 %) es el tercer tumor más diagnosticado en 2018 detrás del de pulmón (11,6 %) y el de mama (11,6 %). En 2019 este tipo de tumor ha sido el más diagnosticado en España según fuentes de la *Red Española de Registros de Cáncer*.

Según la *American Cancer Society* el cáncer colorrectal normalmente tiene una estrecha relación con el estilo de vida (alimentación, peso y ejercicio). Por lo tanto, el sedentarismo y la mala alimentación es una de las principales causas para enfermar.

En este trabajo final de grado se estudiarán los factores asociados a la supervivencia del cáncer colorrectal mediante un método de análisis no párametrico, el método de Kaplan y Meier, y dos métodos de regresión distintos. Primero, se utilizará el modelo paramétrico de vida acelerada, en segundo lugar, se aplicará el método más conocido y usado, el modelo de regresión semiparamétrico de riesgos proporcionales de Cox.

Este estudio de supervivencia se realizará sobre una base de datos que contiene una gran cantidad de infor-

mación de pacientes diagnosticados de cáncer colorrectal en Estados Unidos durante los años 1973 y 2016.

1.2. Cáncer colorrectal

La gran mayoría de información de este apartado ha sido obtenida de la **American Cancer Society** (ver **Cancercolorrectal**,*Factores de riesgo del cáncer colorrectal* [9],*Tratamiento del cáncer de colon según la etapa* [18],*Tasas de supervivencia por etapas para el cáncer colorrectal* [16]), del **Instituto Nacional del Cáncer** (ver *¿Qué es el Cáncer?* [2])y de la **Sociedad Española de Oncología Médica** (ver *Cáncer de colon y recto* [4]).

La palabra cáncer habla de un conjunto de enfermedades relacionadas, en todas ellas, algunas de las células del cuerpo se dividen sin detenerse y se diseminan a los tejidos de alrededor. Puede empezar en casi cualquier parte del cuerpo. El cáncer provoca que un proceso común y ordenado de las células humanas se descontrole y formen masas de células, estos se llaman tumores.

El cáncer de Colon o recto, es el tipo de cáncer que se origina en el colon o en el recto, se agrupan porque tienen muchas características comunes.

Estos tipos de cáncer comienzan como un crecimiento en el revestimiento interno del colon o el recto y son nombrados como pólipos. No todos ellos se convierten en cáncer, los que finalmente lo hacen, suelen tardar unos años. Dependiendo del tipo de pólipo hay una probabilidad distinta de que pueda acabar convirtiéndose en cáncer. Los dos tipos principales son:

- **Pólipos adenomatosos:** tienen una gran probabilidad de convertirse en cáncer, por este hecho se denominan afecciones precancerosas.
- **Pólipos inflamatorios e hiperplásicos:** estos tipos de pólipos son los más comunes, pero tienen unas probabilidades muy reducidas de convertirse en cáncer.

Otros factores que determinan si se podría llegar a desarrollar un cáncer serían:

- Si el pólipo es mayor a un cm.
- Si se descubre que hay más de dos pólipos.
- Si se extirpa un pólipo y se descubre la presencia de *displasia*, la cual es una afección precancerosa.

Normalmente este cáncer se forma en el interior del pólipo y con el tiempo puede crecer hasta la pared del colon o recto, en dicha pared hay muchas capas. Este tipo de cáncer se suele originar en la capa más interna (mucosa) y crece al exterior a través de las demás capas. Una vez están en la pared pueden crecer hacia los vasos linfáticos o sanguíneos, y desde ahí desplazarse a diferentes partes del cuerpo.

Los tipos de cáncer que suelen aparecer en el colon y el recto son los siguientes:

- **Adenocarcinomas:** representan alrededor del 96 % de cánceres colorrectales. Son originados por las células que producen mucosidad para lubricar el interior del colon y el recto.
- **Tumores carcinoides:** estos tipos de tumor son originados a partir de las células productoras de hormonas en el intestino.
- **Tumores estromales gastrointestinales:** las células especializadas de la pared del colon (*Células intersticiales de Caja*) son las que originan este tipo de tumor. Pueden ser encontrados en cualquier parte del sistema digestivo, muy poco común en el colon.
- **Linfomas:** estos tumores provienen de las células del sistema inmunológico, la mayoría son originados en los ganglios linfáticos, pero también hay casos de comienzos en el colon o recto.
- **Sarcomas:** suelen originarse en los vasos sanguíneos, capas musculares u otros tejidos como la pared del colon y recto.

1.2.1. Factores de riesgo

A continuación se detallarán los factores que según datos de la *American Cancer Society* afectan a la probabilidad de padecer pólipos o cáncer colorrectal, estos se dividirán en dos grupos, factores de riesgo que se pueden cambiar y los que no.

- **Factores de riesgo que se pueden cambiar:**
  - **Sobrepeso u obesidad:** las personas que padecen sobrepeso u obesidad tienen una mayor probabilidad de padecer este tipo de cáncer, el impacto suele ser mayor en hombres.
  - **Tabaquismo:** las personas fumadoras tienen una mayor probabilidad de desarrollar cáncer colorrectal respecto a las personas no fumadoras.
  - **Inactividad física:** una persona que no es activa físicamente tiene una mayor probabilidad de desarrollar un cáncer colorrectal.
  - **Ciertos tipos de alimentos:** una dieta a base de carnes rojas y procesadas aumenta la probabilidad de desarrollar este tipo de cáncer, dicha probabilidad también se ve aumentada a causa del consumo de carnes cocinadas a temperaturas muy altas.
  - **Consumo de alcohol:** el consumo excesivo de alcohol está asociado a muchos tipos de cáncer, entre ellos, el colorrectal.
- **Factores de riesgo que no se pueden cambiar:**
  - **Envejecimiento:** Con la edad el riesgo de padecer cáncer en el colon o en el recto es mayor, sobretodo a partir de los 50 años.
  - **Antecedentes personales de:**
    - **Cáncer o pólipo colorrectal:** Las personas que han padecido este cáncer, tienen una mayor probabilidad de recaer en alguna otra parte del colon o del recto.

  - **Enfermedad inflamatoria intestinal (IBD):** Las personas que han padecido alguna enfermedad inflamatoria intestinal tienen un mayor riesgo. En algunos casos que estas enfermedades no se han tratado a tiempo se ha llegado a desarrollar el cáncer.

- **Antecedente familiar de cáncer colorrectal o pólipos adenomatosos:** La existencia de un familiar de primer grado que ha padecido este tipo de cáncer aumenta la posibilidad de padecerlo. Además si dicho familiar lo padeció antes de los 45 años, la probabilidad es aún mayor.
- **Síndromes hereditarios:** alrededor del 5 % de personas que padecen este tipo de cáncer presentan mutaciones que pueden causar síndromes de cáncer familiares y que llevan a padecer la enfermedad.
  - **Síndrome de Lynch:** es el síndrome de cáncer colorrectal hereditario más común, estos tipos de cánceres suelen desarrollarse cuando las personas son jóvenes. Las personas que desarrollan esta mutación (defecto en el gen *MSH2* o *MLH1*) pueden llegar a tener un riesgo del 80 % de padecerlo.
  - **Poliposis adenomatosa familiar (FAP):** El 1 % de cánceres colorrectales son a causa de esta mutación en el gen *APC*.
- **Antecedentes étnicos:** Las personas afroamericanas tienen una de las tasas de incidencia más altas de desarrollar este tipo de tumor en Estados Unidos. A nivel mundial, los judíos procedentes de Europa Oriental son los que tienen el mayor riesgo de padecer este tipo de cáncer.
- **Diabetes tipo 2:** Las personas que padecen la diabetes de tipo 2 tienen un mayor riesgo de padecer este tipo de cáncer. Uno de los motivos, es que ambas enfermedades comparten muchos factores de riesgo. Sin contar estos factores las personas que padecen este tipo de diabetes siguen teniendo una mayor probabilidad de padecer cáncer colorrectal.

1.2.2. Detección y tratamiento

En las últimas décadas la temprana detección de pólipos en el colon y recto ha ayudado a disminuir la tasa de muertes. La razón principal ha sido que en la actualidad las pruebas de detección detectan los pólipos antes de que estos se conviertan en cancerígenos. Un pólipo puede tardar entre 10 y 15 años en convertirse en cáncer.

Si no se consigue extirpar el pólipo antes de convertirse en cáncer pero aun no se ha propagado fuera del colon o recto, se considera como etapa inicial. Esta etapa tiene una supervivencia bastante alta, pero solo el 40 % de los enfermos se detectan en esta etapa. Cuando se supera esta etapa, la tasa de supervivencia es mucho más baja.

Según la *American Cancer Society*, en Estados Unidos un 33 % de las personas que requiere someterse a las pruebas de detección nunca lo ha hecho. Se puede deber a distintos motivos, pero los dos principales son: no saben que hacer estas pruebas de detección de forma periódica podría salvarlos de esta enfermedad o tiene un elevado coste económico si no tienes un seguro médico.

El tratamiento de este tipo de cáncer se basa en la etapa que se encuentra. Dividiremos el tratamiento según en la etapa que se encuentre el tumor:

- **Etapa 0:** Los tumores en etapa 0 no se han propagado más allá del revestimiento interno del colon, normalmente todo lo que se necesita para tratarlo es una cirugía que extrae el cáncer.
- **Etapa I:** En esta etapa el cáncer ya ha crecido más profundamente hacia las capas de la pared, pero no se ha propagado fuera de dicha pared. Puede haber tres casos distintos de cáncer en esta etapa:
  - Si el cáncer ha sido parte de un pólipo y se extrae completamente sin células cancerosas en los bordes, puede que no sea necesario ningún otro tratamiento.
  - Si hay células cancerosas en el borde del polipo, se recomendaría más cirugía. En el caso que no se pueda extraer completamente, dificultará ver si había células cancerosas en el borde, por lo tanto, se podría recomendar más cirugía.
  - Si el cáncer no está en un pólipo, el tratamiento consiste en extirpar la parte del colon que contiene cáncer junto a los ganglios linfáticos cercanos. Generalmente, no hará falta ningún tratamiento más.
- **Etapa II:** En esta etapa el cáncer posiblemente haya crecido a través de la pared del colon y posiblemente a los tejidos cercanos pero aun no se ha extendido a ganglios linfáticos. En algunos casos el único tratamiento que se necesite sea la cirugía para extirpar la sección del colon que contiene cáncer junto a los ganglios linfáticos que se necesite. Sin embargo en otros casos se puede recomendar quimioterapia, principalmente cuando hay un riesgo de recurrencia.
- **Etapa III:** En la etapa III, los tumores se han propagado a los ganglios linfáticos cercanos pero aún no se han extendido a otras partes del cuerpo. Comúnmente, el tratamiento para esta etapa consiste en una cirugía para extirpar el cáncer junto con los ganglios linfáticos cercanos y un tratamiento de quimioterapia. Para el caso de personas que no están lo suficientemente saludables para someterse a cirugía, se les practica la radioterapia, quimioterapia o ambas en conjunto.
- **Etapa IV:** Esta etapa es la más grave, en este momento los cánceres se han propagado hasta los órganos y tejidos cercanos. En muchos casos el cáncer de colon se propaga al hígado, aunque también puede que se propague a otras partes del cuerpo. Normalmente la cirugía no puede curar estos casos, a no ser que haya pocas y pequeñas áreas de propagación en el hígado o pulmones, la cirugía puede ayudar a vivir por más tiempo. Generalmente, también se administra quimioterapia antes y/o después de la operación.

  Si el cáncer se ha propagado demasiado para poder realizar cirugía, el tratamiento principal es la quimioterapia. Si el cáncer bloquea el colon a veces se realiza una cirugía para desbloquearlo.
- **Cáncer recurrente:** Se habla de cáncer recurrente cuando el tumor vuelve a aparecer después de la finalización del tratamiento. Puede ser local (cerca del área del tumor inicial) o en órganos distantes.
  - Recurrencia local: Normalmente la cirugía seguida de quimioterapia puede ayudar a vivir más tiempo y puede incluso curarlo. Si no se puede eliminar con cirugía se puede usar quimioterapia primero, si esta resulta exitosa y el tamaño del tumor disminuye, el uso de la cirugía y más quimioterapia después podría ser una opción a considerar.
  - Recurrencia distante: El tratamiento de este tipo de cáncer suele ser parecido a la recurrencia local.

Generalmente este tipo de cánceres suelen ser más difíciles de tratar.

1.2.3. Mortalidad y supervivencia

Para hablar de la supervivencia de este cáncer se usará la medida de tasa relativa de supervivencia a 5 años, esta tasa compara las personas que tienen el mismo tipo y etapa de cáncer con la población general. El porcentaje de tasa relativa de supervivencia a 5 años querrá decir la probabilidad de supervivencia respecto a las personas que no padecen este cáncer a vivir al menos 5 años después del diagnóstico.

Etapa	**Tasa Relativa de supervivencia a 5 años**
Localizado	90 %
Regional	71 %
Distante	14 %
Todas las etapas SEER combinadas	64 %

Tabla 1.1: Porcentajes de supervivencia cáncer de colon.

Etapa	**Tasa Relativa de supervivencia a 5 años**
Localizado	89 %
Regional	70 %
Distante	15 %
Todas las etapas SEER combinadas	67 %

Tabla 1.2: Porcentajes de supervivencia cáncer de recto.

Los porcentajes nombrado en las tablas 1.1 y 1.2 provienen de la base de datos del *"Programa de Vigilancia, Epidemiología y Resultados"* en inglés se nombra *"Surveillance, Epidemiology, and End Results"* y tiene las siglas SEER, este programa es mantenido por el *"Instituto Nacional del Cáncer "* (NCI). El SEER divide los cánceres en tres etapas:

- **Localizado:** En esta etapa no hay señal de propagación del cáncer fuera del colon o recto.
- **Regional:** Para esta etapa el cáncer se ha propagado hacia las estructuras o ganglios linfáticos más próximos.
- **Distante:** En esta etapa el cáncer ya se ha propagado a partes lejanas como el hígado o pulmones. Este incluye cánceres de etapa IV.

Como era esperable los resultados de supervivencia para los cánceres en etapa "Distante"tienen una supervivencia muy baja, en cambio, la supervivencia para los tipos de tumores en la etapa "Localizado.[es] cercana al 100 %.

1.3. Objetivos

El principal objetivo de este trabajo de fin de grado es analizar los principales factores asociados a la supervivencia de los pacientes diagnosticados con cáncer colorrectal en EEUU recogidos por la base de datos del *Instituto Nacional de Estados Unidos*.

Dicho análisis de supervivencia se analizará de una forma semiparamétrica utilizando el modelo de Cox y de una forma paramétrica con el modelo de Vida Acelerada. Se buscará que modelo se ha ajustado mejor a los datos y que conclusiones se han extraído de los factores asociados a la supervivencia.

El análisis será realizado mediante el software R, se usarán paquetes específicos para el análisis de supervivencia y para los modelos de Cox y Vida Acelerada.

En el Capítulo II se explicará de donde proviene la base de datos, así como fue su lectura en R y las variables que se usarán en el estudio con un breve resumen numérico de estas. Los dos modelos usados en el trabajo estarán descritos detalladamente en el Capítulo III. El Capítulo IV mostrará los resultados que se han obtenido utilizando los distintos modelos. Por último, en el capítulo 5 se describirán las conclusiones del trabajo.

Capítulo II

Base de Datos

2.1. Programa de Vigilancia, Epidemiología y Resultados

El programa de Vigilancia, Epidemiología y Resultados, en inglés con las siglas SEER (Surveillance, Epidemiology and End Results Program), el cual pertenece al Instituto Nacional del Cáncer de Estados Unidos (NCI), es el que se encarga de recoger y publicar los datos de la base de datos que se ha usado en este trabajo.

Los datos de la incidencia del cáncer recogidos por SEER recogen aproximadamente el 34.6 % de los registros poblacionales de cáncer en EEUU. Se recoge información como procedencia o etnia del paciente, lugar del tumor primario, estado del tumor en el momento del diagnóstico. Más adelante se detallará mejor las variables de esta base de datos que han sido recogidas y utilizadas en el posterior análisis de supervivencia. El sitio web para obtener más información y solicitar acceso a la base de datos es el siguiente (http://seer.cancer.gov/data). Para poder usar la base de datos en este trabajo, se tuvo que pedir un tipo de permiso para uso académico.

Además de recoger estos datos, SEER también trabaja en la publicación de distintos artículos académicos relacionados con el cáncer, el más importante es el Informe Anual sobre el Estado del Cáncer, es elaborado gracias a la colaboración entre expertos del Instituto Nacional del Cáncer (NCI), Centros para el Control y Prevención de Enfermedades (CDC), la Sociedad Americana del Cáncer (ACS) y la Asociación de Registros Centrales del Cáncer de América del Norte (NAACCR).

2.2. Lectura de datos

Una vez obtenido el permiso, SEER envió un nombre de usuario y una contraseña para poder descargar una carpeta en formato zip que contenía los datos en formato ASCII y distintos documentos de explicación de estos. En mi caso, descargué los datos que provenían desde 1975 hasta 2016. Estos datos, tenían un archivo txt para cada tipo de cáncer (pulmón, colorrectal, genitales masculinos, genitales femeninos, linfoma-melanoma-leucemia, otro tipo de digestivos,respiratorio, tracto urinario y otros), usé el de cáncer colorrectal.

En el documento txt encontramos que cada fila del documento es un registro, mientras que las variables no

tienen ningún elemento de separación. Por lo tanto, para poder definir correctamente que datos correspondían a cada variable, fue necesario consultar el documento *"TextData.FileDescription.pdf"* incluido en la carpeta comprimida. En este documento se pudo encontrar cuál era la columna inicial y final de cada variable, con esta información y el uso de la función *read_fwf* contenida en el paquete Wickham, Hester y Francois [20] [20] *readr* se pudo leer correctamente la base de datos en *R*. El código de R correspondiente se muestra en el Código 1.

```
if (!(require(readr))){
  install.packages("readr")
}
data <- read_fwf(file=path_dades_1975_2016,col_positions=fwf_positions(start=
    start_col,end=end_col,col_names=colname_dades_1975_2016))
```

Código 1: Lectura de la base de datos en R.

La base de datos completa contiene un total de 551174 registros y 142 variables, en nuestro caso se usarán 23 variables. A través del paquete *usmap* (ver Di Lorenzo [7]) y la variable *Registry ID*, se ha dibujado un gráfico de Estados Unidos con la distribución de cantidad de datos por estados. En el Código 2 se puede ver el código de R usado para la creación del mapa.

```
data("statepop")
statepop <- statepop$abbr[statepop$abbr!="CT"]
statepop <- statepop[statepop!="HI"]
statepop <- statepop[statepop!="IA"]
statepop <- statepop[statepop!="GA"]
statepop <- statepop[statepop!="MI"]
statepop <- statepop[statepop!="NM"]
statepop <- statepop[statepop!="CA"]
statepop <- statepop[statepop!="WA"]
statepop <- statepop[statepop!="UT"]
mapdata <- data.frame("state"=c("CT","HI","IA","GA","MI","NM","CA","WA","UT",
    statepop),"SEER_NAMES"=c(names(table(dades$REGISTRY_ID)),rep(NA,length(statepop
    ))),count=c(as.vector(table(dades$REGISTRY_ID)),rep(0,length(statepop))))

plot_usmap(data = mapdata, values = "count", color = "black",labels=T) +
scale_fill_continuous(low = "Blanco", high = "blue",name = "Casos en base de datos
    Total", label = scales::comma) +  theme(legend.position = "right")
```

Código 2: Creación mapa de Estados Unidos en R para comprobar la distribución de los datos.

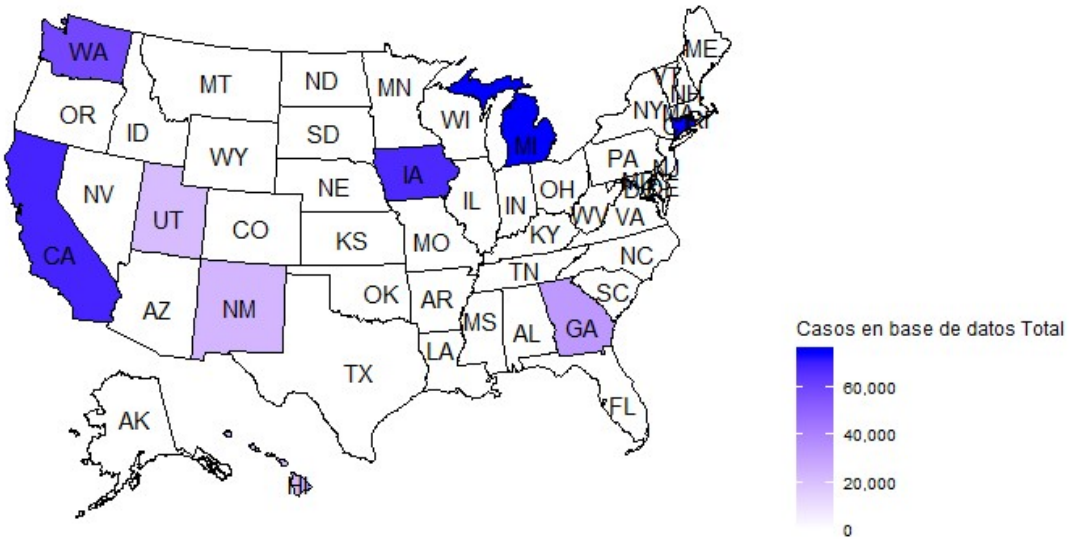


Figura 2.1: Mapa de Estados Unidos con la distribución de los datos.

Se puede observar en el Gráfico 2.1 como no se han recogido datos de todos los estados del país, se han recogido datos de 9 estados distintos, a pesar de esto, se puede ver en el gráfico que se ha intentado abarcar las distintas zonas del país, e incluso se han recogido datos de una isla perteneciente a Estados Unidos (Hawaii).

2.3. Definición de variables usadas

A continuación, se presentarán en la Tabla 2.1 las variables que se han usado en la primera exploración de los datos, su significado y el número y porcentaje de *missings* de cada una.

Variable	Descripción	Missing (%)
ID del paciente	Código único asignado a cada paciente.	0 (0 %)
ID del registro	Código que identifica la población del paciente. Se ha descodificado para tener directamente el nombre de la población.	0 (0 %)
Estado civil	Estado civil del paciente en el momento del diagnóstico.	0 (0 %)
Etnia	Etnia del paciente.	0 (0 %)
Género	Género del paciente.	0 (0 %)
Edad	Edad del paciente en el momento del diagnóstico.	17 (0.0038 %)
Año de diagnóstico	Año en el que se diagnostica el cáncer.	0 (0 %)
Zona primaria	Lugar donde se originó el primer tumor.	0 (0 %)
Grado	Estado del tumor en el momento del diagnóstico.	0 (0 %)
Confirmación del diagnóstico	Identifica el método usado para confirmar el tumor.	0 (0 %)
Fuente de datos	Lugar donde se han obtenido los datos.	0 (0 %)
Nodos positivos	Número de nodos con metástasis alrededor del tumor.	227170 (50.82 %)
Tamaño del tumor	Tamaño del tumor en mm.	353822 (79.16 %)
Cirugía	Razón por no realizar la cirugía o si esta ha sido realizada.	0 (0 %)
Tipo de tumor	Identifica el tipo de tumor.	0 (0 %)
Indicador primario de tumor maligno	Identifica si en un primer momento el tumor se identificó como maligno o no.	0 (0 %)
Estado vital	Identifica si la persona está viva o muerta.	0 (0 %)
Causa específica de muerte	Identifica si la persona ha muerto de cáncer o por otro motivo.	0 (0 %)

Variable	Descripción	Missing (%)
Meses de supervivencia	Meses que ha sobrevivido el paciente al tumor.	0 (0 %)
Seguro médico	Seguro médico que tenia el paciente en el momento del diagnóstico.	343923 (76.95 %)

Tabla 2.1: Descripción de variables usadas y número y porcentaje de *missings*.

Una vez hecha la primera exploración de los datos, se detectó que las variables **Tamaño del tumor** y **Seguro médico** tenían un gran porcentaje de *missings*. Al detectarlo, se procedió a explorar el comportamiento de dichas variables y se encontró que no se recogieron datos de estas dos variables hasta fechas recientes, en el caso de **Tamaño del tumor** se empezaron a recoger datos a partir del 2004 y para la variable **Seguro médico** se recogieron datos a partir del 2007. Ante esta problemática, se decidió dividir la base de datos principal en dos bases de datos más pequeñas a través de la variable **Año de diagnóstico**. La primera base de datos contemplará los datos desde 1975 hasta 1984 y la segunda irá desde 2004 hasta 2011. Los datos no podrán recogerse mas allá de 2011, ya que se comprobará la supervivencia hasta los 5 años y estos datos contienen información hasta 2016.

Finalmente, la base de datos del período de 1975 hasta 1984 contiene 98527 observaciones y 19 variables, la base de datos del período de 2004 hasta 2011 contiene 84683 observaciones y 22 variables.

2.4. Resumen descriptivo univariante de variables usadas

En este apartado se pretenderá explorar la distribución de la información de cada variable, se estudiarán las que existe intención de usar en el posterior análisis de supervivencia.

2.4.1. Variables numéricas

Para empezar, se explorarán las variables continuas para las dos bases de datos. Los gráficos se realizarán con el paquete de R *ggplot* (ver Wickham [19]).

Período 1975-1984

En la Tabla 2.2 se presentará un análisis univariante de la variable **Edad**, para los pacientes que han sido diagnosticados con cáncer colorrectal entre 1975 y 1984.

Variable	Media	Mediana	Desv. est.	Min.	Max.
Edad [Años]	68.63	69	12.35	4	106

Tabla 2.2: Descriptiva de la variable numérica Edad período 1975-1984.

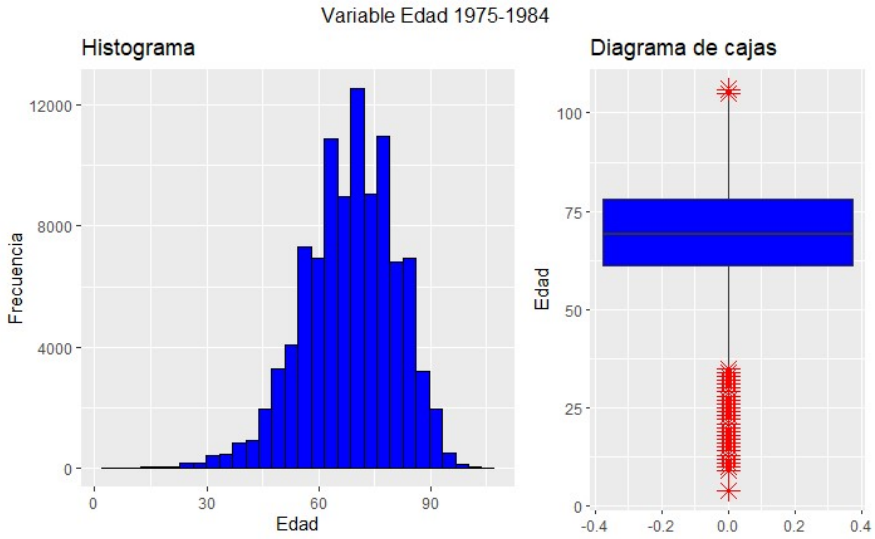


Figura 2.2: Histograma y diagrama de cajas para la variable **Edad** de la base de datos del intervalo de años entre 1975-1984 en R.

Como es observable en el Gráfico 2.2, la distribución de la variable **Edad** es bastante simétrica con un ligero desplazamiento a la derecha. Se puede ver como esta variable comprende una gran cantidad de años (desde 4 hasta 106), pero la gran cantidad pacientes que se estudiarán tenían una edad entre los 60 y 78 años en el momento del diagnóstico, se puede ver en el diagrama de cajas del Gráfico 2.2.

Período 2004-2011

En la Tabla 2.3 se presentará la información numérica que se ha considerado más relevante para las variables numéricas de la base de datos que corresponde a los pacientes que han sido diagnosticados con cáncer colorrectal entre 2004 y 2011.

Variable	Media	Mediana	Desv. est.	Min.	Max.
Edad [Años]	66.5	67	14.33	0	108
Número de nodos positivos	1.86	0	3.89	0	80
Tamaño del tumor [mm]	45.6	40	34.32	0	989

Tabla 2.3: Descriptiva de variables numéricas del intervalo de años entre 2004 y 2011.

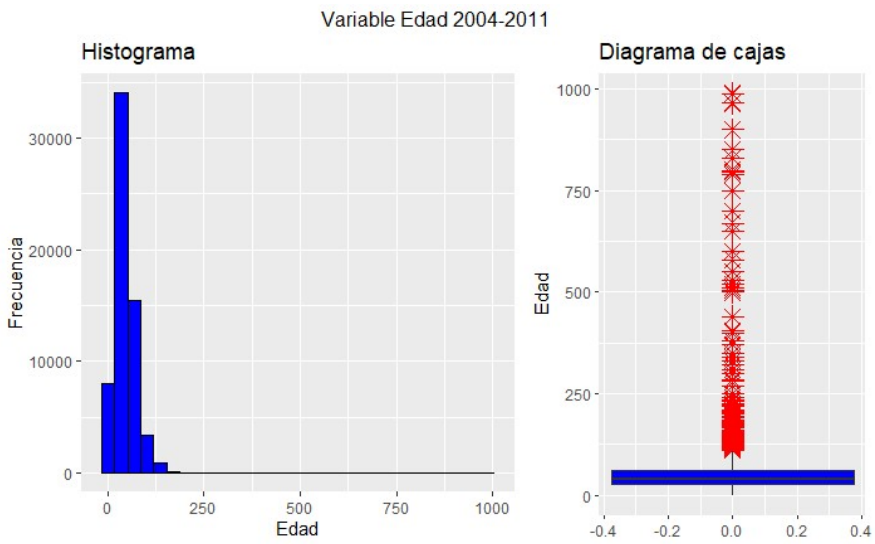


Figura 2.3: Histograma y diagrama de cajas para la variable **Tamaño del tumor** de la base de datos del intervalo de años entre 2004-2011 en R.

Se puede observar en el Gráfico 2.3, como la distribución de la variable **Edad** para esta base de datos, es similar a la de la base de datos de los años 1975-1984. La variable de **Nodos positivos** se descartará ya que se ha visto en los análisis posteriores que no aporta información.

Como también se puede ver de forma gráfica, para la variable correspondiente al **Tamaño del tumor**, la gran mayoría de los datos tienen valores menores a 80 (ver diagrama de cajas del Gráfico 2.3), a pesar de esto, existe una gran cantidad de datos extremos que llegan hasta 989. Esto hace que se distribuya de una forma asimétrica totalmente desplazada hacia la izquierda.

2.4.2. Variables categóricas

En este apartado se procederá a explorar numéricamente y gráficamente las variables categóricas de las dos bases de datos.

Período 1975-1984

En la Tabla 2.4 se presentarán las variables categóricas de la base de datos correspondiente al intervalo de tiempo entre 1975 y 1984, las categorías de cada una y su frecuencia y porcentaje.

Variable	Categorías	Frecuencia (%)
Estado civil	Soltero	7209 (7.32 %)
	Casado	57094(57.95 %)
	Separado	4966 (5.04 %)
	Divorciado	4112(4.17 %)
	Viudo	22317 (22.65 %)
	No casado	0 (0 %)
	Desconocido	2829 (2.87 %)
Etnia	Blanca	87588 (88.9 %)
	Afroamericana	6805 (6.91 %)
	Asiática o isleña	3825 (3.88 %)
	Otras	309 (0.31 %)
Género	Hombre	49112 (49.85 %)
	Mujer	49415 (50.15 %)
Grado	Grado I	14331 (14.55 %)
	Grado II	34783(35.3 %)
	Grado III	12375(12.56 %)
	Grado IV	1322 (1.34 %)
	Desconocido	35716 (36.25 %)
Confirmación del diagnóstico	Histológica	94194 (95.6 %)
	Citológica	130 (0.13 %)
	Histológica	0 (0 %)
	Microscópica	35 (0.04 %)
	Laboratorio	0 (0 %)
	Visualización directa microsc	909 (0.92 %)
	Radiología	2236 (2.27 %)
	Diagnóstico clínico	754 (0.77 %)
	Desconocido	269 (0.27 %)
Fuente de datos	Hospital del paciente	97747 (99.21 %)
	Radioterapia	0 (0 %)
	Laboratorio	162 (0.16 %)
	Medico privado	594 (0.6 %)
	Enfermería	24 (0.02 %)
	Otro hospital	0 (0 %)
Cirugía	Cirugía realizada	85699 (86.98 %)
	No recomendada	0 (0 %)
	Contraindicada	0 (0 %)
	Paciente murió antes	0 (0 %)
	Paciente rechazó	0 (0 %)
	Desconocido	12828 (13.02 %)

Variable	Categorías	Frecuencia (%)
Tipo de tumor	Maligno	94973 (96.39 %)
	No maligno	3554 (3.61 %)
Indicador primario de tumor maligno	Sí	94973 (96.39 %)
	No	3554 (3.61 %)
Estado vital	Muerto	95140 (96.56 %)
	Vivo	3387 (3.44 %)

Tabla 2.4: Descriptiva de variables categóricas para la base de datos 1975-1984.

Viendo los resultados del análisis numérico, se ha llegado a la conclusión que las variables **Cirugía**, **Tipo de tumor**, **Indicador primario de tumor maligno**, **Estado vital**, **Fuente de Datos** y **Confirmación del diagnóstico**, no se usarán en el análisis de supervivencia, ya que casi todos los datos se concentran en una categoría, por lo tanto, no se obtendrá información interesante.

En el Gráfico 2.4 se puede observar la distribución de las variables **Grado**, **Etnia** y **Estado Civil**.

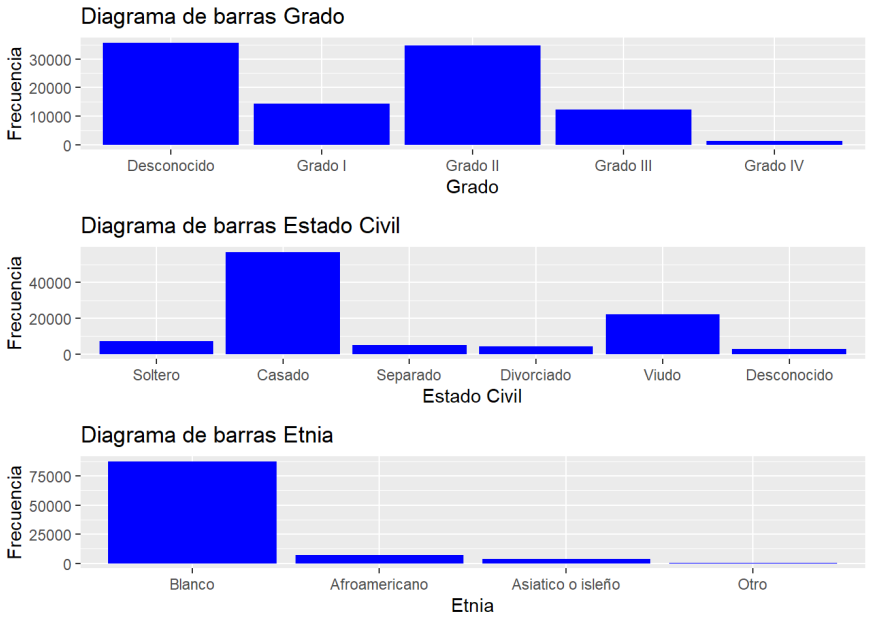


Figura 2.4: Diagrama de barras para las variables **Etnia**, **Grado** y **Estado Civil** de la base de datos del intervalo de años entre 1975-1984 en R.

Se puede comprobar como las categorías predominantes son *Grado II* para la variable **Grado**, *Casado* para **Estado Civil** y *Blanca* en la variable **Etnia**.

Período 2004-2011

En la Tabla 2.5 se presentarán las categorías de cada variable categórica, su frecuencia y porcentaje para los pacientes correspondientes al intervalo de tiempo entre 2004 y 2011.

Variable	Categorías	Frecuencia (%)
Estado civil	Soltero	11556 (13.65 %)
	Casado	44493 (52.54 %)
	Separado	625 (0.74 %)
	Divorciado	7691 (9.08 %)
	Viudo	14864 (17.55 %)
	No casado	18 (0.02 %)
	Desconocido	5436 (6.42 %)
Género	Hombre	43316 (51.15 %)
	Mujer	41367 (48.85 %)
Etnia	Blanca	65330 (77.15 %)
	Afroamericana	9651 (11.4 %)
	Asiática o isleña	8430 (9.95 %)
	Otras	1272 (1.5 %)
Grado	Grado I	7223 (8.53 %)
	Grado II	46761 (55.22 %)
	Grado III	12549 (14.82 %)
	Grado IV	1545 (1.82 %)
	Desconocido	16605 (19.61 %)
Confirmación del diagnóstico	Histológica	82599 (97.54 %)
	Citológica	233 (0.28 %)
	Histológica	0 (0 %)
	Microscópica	18 (0.02 %)
	Laboratorio	28 (0.03 %)
	Visualización directa microsc	146 (0.17 %)
	Radiología	1065 (1.26 %)
	Diagnóstico clínico	280 (0.33 %)
	Desconocido	314 (0.37 %)
Fuente de datos	Hospital del paciente	80666 (95.26 %)
	Radioterapia	277 (0.33 %)
	Laboratorio	1638 (1.93 %)
	Medico privado	1334 (1.58 %)
	Enfermería	61 (0.07 %)
	Otro hospital	707 (0.83 %)

Variable	Categorías	Frecuencia (%)
Cirugía	Cirugía realizada	71955 (84.97 %)
	No recomendada	8502 (10.04 %)
	Contraindicada	1052 (1.24 %)
	Paciente murió antes	138 (0.16 %)
	Paciente rechazó	1051 (1.24 %)
	Desconocido	1985 (2.34 %)
Tipo de tumor	Maligno	81576 (96.33 %)
	No maligno	3107 (3.67 %)
Indicador primario de tumor maligno	Sí	81576 (96.33 %)
	No	3107 (3.67 %)
Estado vital	Muerto	45030 (53.17 %)
	Vivo	39653 (46.83 %)
Seguro médico	No asegurado	1210 (2.32 %)
	Seguro público	4734 (9.07 %)
	Asegurado	43458 (83.27 %)
	Desconocido	2787 (5.34 %)

Tabla 2.5: Descriptiva de variables categóricas de la base de datos 2004-2011.

En la base de datos comprendida entre los años 2004 y 2011, también se eliminarán del posterior análisis las variables que se han eliminado en la base de datos que comprende el intervalo de tiempo entre 1975 y 1984. La variable **Estado Vital** no se ha creído que pudiera ser relevante para el posterior estudio, por lo tanto, se eliminará de la base de datos.

Las variables que se han graficado en el Gráfico 2.4, se distribuyen de una forma muy parecida aunque con porcentajes ligeramente diferentes. A continuación, en el Gráfico 2.5 se podrá ver un diagrama de barras para la variable *Seguro Médico*, la cual no está en la base de datos anterior.

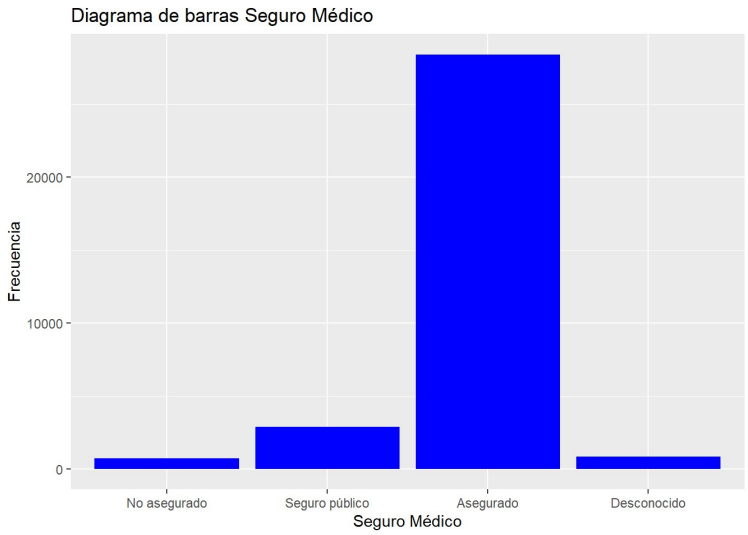


Figura 2.5: Diagrama de barras para las variable **Seguro Médico** de la base de datos del intervalo de años entre 2004-2011 en R.

Esta variable tiene un alto número de *missings*, ya que como se ha comentado anteriormente, se empezó a recoger información a partir de 2007. A pesar de esto, se puede comprobar como la gran mayoría de pacientes que se han estudiado, tenían seguro médico.

2.5. Resumen descriptivo bivariante de variables usadas

Para explorar las interacciones entre las variables que se iban a usar en los análisis de supervivencia se hicieron distintos gráficos y tablas bivariantes en R. Los gráficos se hicieron con el paquete mencionado anteriormente (*ggplot*), para las tablas se usó el paquete *compareGroups* (ver Subirana, Sanz y Vila [15]).

Período 1975-1984

A continuación se mostrará una tabla que se hizo con la función *createTable* del paquete mencionado anteriormente (*compareGroups*) entre la variable continua **Edad** y la variable categórica **Grado**.

	Grado I N=14331	Grado II N=34783	Grado III N=12375	Grado IV N=1322	Desconocido N=35713
Edad	68.2 (11.9)	68.3 (11.9)	68.4 (12.5)	68.0 (12.8)	69.2 (12.9)

Tabla 2.6: Descriptiva bivariante de la variable **Edad** y **Grado** de la base de datos de 1975-1984.

Se puede observar en la Tabla 2.6 como la **Edad** no depende del grado, ya que la media y desviación típica de la variable edad para cada categoría de la variable **Grado** es prácticamente igual.

Período 2004-2011

A continuación, en el Gráfico 2.6 se presentará un gráfico bivariante entre las variables **Etnia** y **Seguro Médico**. Se quiso ver si había alguna diferencia dependiendo de la etnia, aunque como se ha visto anteriormente, la mayoría de pacientes tenían seguro médico.

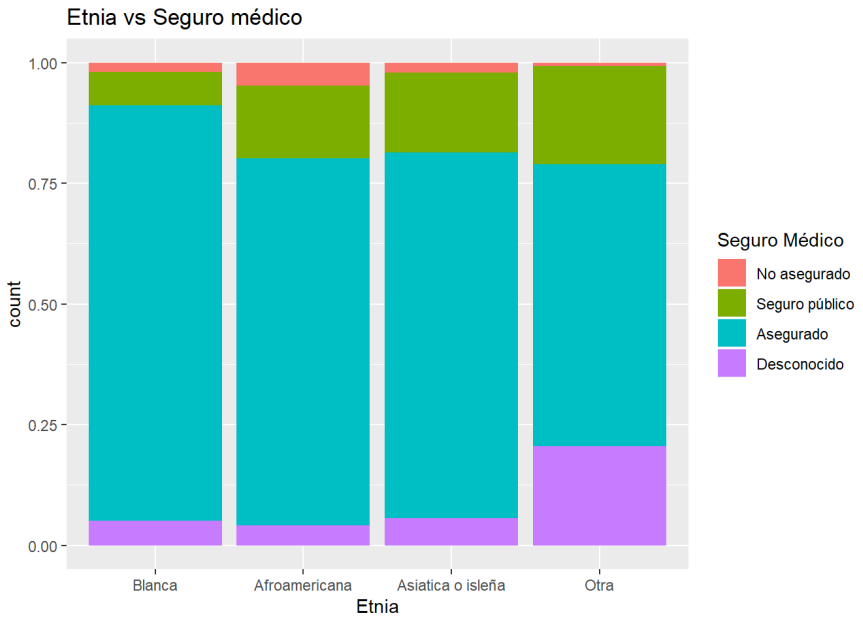


Figura 2.6: Gráfico bivariante para las variable **Seguro Médico** y **Etnia** de la base de datos del intervalo de años entre 2004-2011 en R.

Se puede apreciar como las personas afroamericanas son las que tienen una menor proporción de personas aseguradas respecto a las demás etnias. Otro hecho destacable, es que la categoría *Blanca* tienen la proporción más alta de pacientes con seguro médico, esto es debido a que la proporción de la categoría *Seguro público* es notablemente más baja que en las categorías *Afroamericana* y *Asiática o isleña*, además como se ha mencionado anteriormente es la que tiene una menor proporción de personas no aseguradas.

Capítulo III

Análisis de Supervivencia

En este capítulo se explicarán métodos del ámbito del Análisis de Supervivencia, estos serán usados para analizar los tiempos de supervivencia de los pacientes de cáncer colorrectal.

En el análisis Supervivencia el modelo basado en la aceptación de riesgos proporcionales (modelo de Cox) es el más usado. La principal restricción de este modelo es la asunción de que los coeficientes de regresión son constantes en el tiempo. Una alternativa a este modelo es el modelo de Vida Acelerada, el cual está basado en una distribución paramétrica y se puede comparar con el de Cox cuando sigue una distribución Weibull.

Parte de la elaboración de este capítulo sigue el libro **Análisis de Supervivéncia** de Gómez, Julià y Langohr [10].

3.1. Conceptos básicos y censura

3.1.1. Conceptos básicos

Cuando se habla de análisis de supervivencia hay que tener en cuenta dos factores principales, el tiempo ($\boldsymbol{T}$) y el suceso de interés ($\boldsymbol{\varepsilon}$).

- $\boldsymbol{T}$: debe ser una variable aleatoria no negativa que corresponde a una población homogénea. Normalmente será continua a pesar de que a veces se usan variables aleatorias discretas. En nuestro caso la variable de tiempo son los meses de supervivencia de pacientes de cáncer colorrectal desde el momento del diagnóstico de la enfermedad. Como se ha explicado anteriormente, todos los individuos que superen los 5 años o 60 meses de supervivencia, se consideraran como curados.
- $\boldsymbol{\varepsilon}$: siempre será el objeto de estudio, depende de la investigación puede ser la aparición de un tumor, la muerte, etc. En este trabajo, el suceso de interés es la muerte a causa del cáncer colorrectal, las muertes por otras causas se tratarán como tiempos censurados (ver Sección 3.1.3 en la página 34).

A continuación se presentarán algunas funciones que pueden caracterizar **T**:

- **Función de supervivencia *S(t)*:** corresponde a la probabilidad de que el suceso ε ocurra después del tiempo t, está definida para $t \geq 0$. Siempre será una función monótona decreciente y cuando **T** sea continua, *S(t)* será continua y estrictamente decreciente. La función de supervivencia empezará en 1 y convergerá a 0 cuando t tienda a ∞ (*S(0)*=1 y S(∞)=0).

- **Función de distribución *F(t)*:** es la función complementaria a la función de supervivencia, por lo tanto *F(t)* = 1 - *S(t)*. Al contrario que la función de supervivencia esta empezará en el 0 y convergerá en 1 cuando t tienda a ∞, en consecuencia, será una función monótona creciente.

- **Función de densidad de probabilidad *f(t)*:** esta función es definida para variables completamente continuas y formalmente es definida como el siguiente límite :

$$f(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} Prob[t \leq T < t + \Delta t]$$

  El producto $f(t)\Delta t$ se puede interpretar como la probabilidad de ocurrencia del evento ε en el intervalo $(t,t + \Delta t)$. La función de supervivencia equivaldrá a la integral de la función de densidad $\int_t^\infty f(u)du$.

3.1.2. Función de riesgo

La función de riesgo es la que explica el comportamiento de la probabilidad condicionada de morir por cáncer de colon o recto en un estrecho intervalo de tiempo sabiendo que el individuo empieza estando vivo. Cuando T es continua la función de riesgo es definida como :

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

El producto $\lambda(t)\Delta t$ se puede interpretar como la probabilidad de que a una persona de edad t le ocurra el suceso ε durante el intervalo $(t, t + \Delta t)$. Cuando T es una variable discreta, la función de riesgo expresa la probabilidad de que una persona viva en t_{j-1} muera en t_j, la función se define así:

$$\lambda(t_j) = Prob[T = t_j | T \geq t_j] = Prob[T = t_j | T > t_{j-1}]$$

En el caso que el riesgo instantáneo sea elevado, la función de supervivencia decaerá más rápidamente, en cambio cuando el riesgo sea 0, la curva de supervivencia será plana.

Cuando T es continua:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}(lnS(t))$$

En relación a la función de riesgo, existe la función de riesgo acumulado (Λ). Técnica y gráficamente es muy útil, pero carece de una interpretación intuitiva. Se define de la siguiente manera:

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

Cuando los datos son continuos se cumple la siguiente expresión:

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

3.1.3. Censura

Un obstáculo que se encuentra en el análisis de supervivencia es que la información para la variable de supervivencia de algunos individuos es incompleta. Hay diferentes casuísticas, todas ellas son conocidas como **censura**. En este estudio puede haber tres tipos de casos de censura:

- La persona ha abandonado el estudio antes de los 5 años desde el diagnóstico.
- La persona ha sido diagnosticada a partir del 2012. Existen datos hasta 2016, por lo tanto no se tendría un seguimiento de 5 años. Para evitarlo no se han cogido datos más allá de 2011.
- La persona ha muerto durante el seguimiento por una causa ajena al tumor.

Los 3 casos de censura anteriores, serían casos censurados por la derecha, en concreto de tipo III o censura aleatoria. Siendo $\boldsymbol{T_{(1)}, ..., T_{(n)}}$ los tiempos potenciales hasta el suceso de interés de nuestra muestra de $\boldsymbol{n}$ individuos. Si en el estudio hay censura por la derecha, se procederá a observar los pares $\boldsymbol{(Y_{(1)}, \delta_{(1)}), ..., (Y_{(n)}, \delta_{(n)})}$, donde $\boldsymbol{Y_i = min\{T_i, C_i\}}$, siendo $\boldsymbol{C_1, ...C_N}$ los tiempos de censura para cada individuo $\boldsymbol{y}$. La variable indicadora del evento está definido de la siguiente manera:

$$\boldsymbol{\delta_i = \begin{cases} 1 \quad \text{si} \quad T_i \leq C_i : & \text{el individuo i no esta censurado} \\ 0 \quad \text{si} \quad T_i > C_i : & \text{el individuo i esta censurado} \end{cases}}$$

Se conocen otros tipos de censura que no se repasaran en este trabajo, por ejemplo, la censura por la izquierda, la censura en un intervalo y la censura doble.

3.2. Inferencia no paramétrica

3.2.1. Estimador de Kaplan-Meier

Para la estimación de la función de supervivencia en el análisis no paramétrico de este trabajo se ha usado el método de Kaplan-Meier a través del paquete de R *survival* (ver Therneau [17]). Las primeras fórmulas de este procedimiento estadístico fueron publicadas por Edward L. Kaplan y Paul Meier en 1958 en el articulo Kaplan y Meier [12]. Como se comenta en el artículo Dudley, Wickham y Coombs [8], este método es el más usado en ensayos clínicos aleatorios cuando se cumplen las siguientes premisas:

- Los tratamientos de los pacientes han sido asignados aleatoriamente.
- No todos los pacientes entraron al estudio en el mismo momento

- Pacientes se fueron del estudio o se perdieron en diferentes intervalos de tiempo.
- La variable de interés puede no haber ocurrido durante el período de tiempo.

Los datos utilizados en este trabajo no se basan en un ensayo clínico, por lo tanto, no se ha medido el tratamiento aplicado.

El estimador de Kaplan y Meier (K-M) se mide de forma distinta cuando existen empates en los datos o no. En este estudio hay datos con empates.

Estimador K-M cuando hay empates

Siguiendo la formulación del estimador Kaplan-Meier cuando no hay empates:

$$\hat{\boldsymbol{S}}(\boldsymbol{t}) = \prod_{\boldsymbol{i}:\boldsymbol{Y}_{(i)} \leq \boldsymbol{t}} \hat{\boldsymbol{p}}_{\boldsymbol{i}} = \prod_{\boldsymbol{i}:\boldsymbol{Y}_{(i)} \leq \boldsymbol{t}} \left(\mathbf{1} - \frac{\mathbf{1}}{\boldsymbol{n}_{\boldsymbol{i}}}\right)^{\boldsymbol{\delta}_{(i)}}$$

Se supondrá que después del momento $\boldsymbol{\tau}_{\boldsymbol{k}}$ hay $\boldsymbol{n}_{\boldsymbol{k}}$ individuos vivos y en el momento $\boldsymbol{\tau}_{\boldsymbol{k}}$ se producen $\boldsymbol{d}_{\boldsymbol{k}} > \mathbf{1}$ muertes, se imaginará que se subdivide el intervalo $\boldsymbol{I}_{\boldsymbol{k}}$ en $\boldsymbol{d}_{\boldsymbol{k}}$ intervalos infinitesimales, por lo tanto:

$$\begin{aligned}
\hat{\boldsymbol{p}}_{\boldsymbol{k}} &= \left(\mathbf{1} - \frac{\mathbf{1}}{\boldsymbol{n}_{\boldsymbol{k}}}\right) \cdot \left(\mathbf{1} - \frac{\mathbf{1}}{\boldsymbol{n}_{\boldsymbol{k}} - \mathbf{1}}\right) ... \left(\mathbf{1} - \frac{\mathbf{1}}{\boldsymbol{n}_{\boldsymbol{k}} - \boldsymbol{d}_{\boldsymbol{k}} + \mathbf{1}}\right) \\
&= \left(\frac{\boldsymbol{n}_{\boldsymbol{k}} - \mathbf{1}}{\boldsymbol{n}_{\boldsymbol{k}}}\right) \cdot \left(\frac{\boldsymbol{n}_{\boldsymbol{k}} - \mathbf{2}}{\boldsymbol{n}_{\boldsymbol{k}} - \mathbf{1}}\right) ... \left(\frac{\boldsymbol{n}_{\boldsymbol{k}} - \boldsymbol{d}_{\boldsymbol{k}}}{\boldsymbol{n}_{\boldsymbol{k}} - \boldsymbol{d}_{\boldsymbol{k}} + \mathbf{1}}\right) \\
&= \frac{\boldsymbol{n}_{\boldsymbol{k}} - \boldsymbol{d}_{\boldsymbol{k}}}{\boldsymbol{n}_{\boldsymbol{k}}} = \mathbf{1} - \frac{\boldsymbol{d}_{\boldsymbol{k}}}{\boldsymbol{n}_{\boldsymbol{k}}}
\end{aligned}$$

Este estimador admite la siguiente forma general para todo $\boldsymbol{t}$ en el rango que existen datos.

$$\hat{\boldsymbol{S}}(\boldsymbol{t}) = \begin{cases} \mathbf{1} & \textbf{si} \quad \boldsymbol{t} < \boldsymbol{Y}_{(1)} \\ \prod_{\boldsymbol{i}:\boldsymbol{Y}_{(i)} \leq \boldsymbol{t}} \left(\mathbf{1} - \frac{\boldsymbol{d}_{\boldsymbol{i}}}{\boldsymbol{n}_{\boldsymbol{i}}}\right) & \textbf{si} \quad \boldsymbol{t} \geq \boldsymbol{Y}_{(1)} \end{cases}$$

Si la última observación ordenada está censurada se redefinirá la función de supervivencia $\hat{\boldsymbol{S}}(\boldsymbol{t})$.

La varianza del estimador de Kaplan y Meier viene dada por la fórmula de Greenwood:

$$\hat{\boldsymbol{V}}_{\boldsymbol{G}}(\hat{\boldsymbol{S}}(\boldsymbol{t})) = \hat{\boldsymbol{S}}(\boldsymbol{t})^{\mathbf{2}} \sum_{\boldsymbol{i}:\boldsymbol{Y}_{(i)} \leq \boldsymbol{t}} \frac{\boldsymbol{d}_{\boldsymbol{i}}}{\boldsymbol{n}_{\boldsymbol{i}}(\boldsymbol{n}_{\boldsymbol{i}} - \boldsymbol{d}_{\boldsymbol{i}})}$$

Estimación de la función de riesgo y riesgo acumulada

La función de riesgo para el momento $\boldsymbol{Y}_{(\boldsymbol{i})}$ se estima a partir del número de eventos en este momento dividido por el número de elementos a riesgo en $\boldsymbol{Y}_{(\boldsymbol{i})}$. Por lo tanto, $\hat{\boldsymbol{\alpha}}(\boldsymbol{t}) = \frac{\boldsymbol{d}_{\boldsymbol{i}}}{\boldsymbol{n}_{\boldsymbol{i}}}$.

El estimador general propuesto por Nelson primero y luego por Aalen para la función de riesgo es:

$$\hat{\alpha}_{NA}(t) = \begin{cases} 0 & \text{si} \quad t \neq Y_{(1)} \\ \frac{d_i}{n_i} & \text{si} \quad t = Y_{(1)} \end{cases}$$

3.2.2. Intervalos de confianza

Función de supervivencia

Los intervalos de confianza correspondientes a la función de supervivencia que han sido calculados en diversos tiempos de interés t nos indican cuán fiable es el estimador en cada punto. El intervalo de confianza para la función de supervivencia en el momento inicial (t_0) se denotará así:

$$I_{lineal(t_0)} = \left[\hat{S}(t) - Z_{1-\alpha/2}\sigma_S(t_0)\hat{S}(t_0), \hat{S}(t) + Z_{1-\alpha/2}\sigma_S(t_0)\hat{S}(t_0)\right]$$

Denotando $\sigma^2_S(t) = \frac{\hat{V}_G(\hat{S}(t))}{(\hat{S}(t))^2}$

Función de riesgo acumulada

Se obtienen de forma parecida a los de la función de supervivencia, los intervalos para el momento t_0 serán:

$$I_{lineal(t_0)} = \left[\hat{\Lambda}(t_0) - Z_{1-\alpha/2}\sigma_{NA}(t_0), \hat{\Lambda}(t_0) + Z_{1-\alpha/2}\sigma_{NA}(t_0)\right]$$

Denotando $\sigma_{NA}(t) = \sqrt{\sum_{i:Y_{(i)}\leq t} \frac{d_i}{n_i^2}}$

3.2.3. Prueba del logrank

La prueba del logrank también es conocida como la de Cox y Mantel. Esta prueba detecta las alternativas en las que las funciones de riesgo son proporcionales, por lo tanto, corresponderían a funciones de supervivencia que satisfarían la siguiente igualdad : $S_j(t) = S(t)^{\theta_j}$.

La prueba de hipótesis se podría formular de la siguiente forma:

$$H_0 : \lambda_1(t) = \lambda_2(t) \text{ para todo } t \leq \tau \quad \text{vs} \quad H_{RP} : \lambda_2(t) = \exp\beta\lambda_1(t)$$

Si no se obtiene suficiente evidencia para rechazar la hipótesis nula, se dirá que las dos poblaciones tienen la misma función de riesgo. (ver Bland y Altman [3]).

3.3. Modelo de vida acelerada

La prueba logrank permite comparar las funciones de supervivencia de dos o más poblaciones, no obstante a menudo es interesante el estudio de la relación entre $\boldsymbol{T}$ y varias variables numéricas y categóricas. En estos casos, se pueden usar modelos como el de vida acelerada o el de riesgos proporcionales (Cox).

Cuando se supone un modelo paramétrico, los estimadores normalmente se obtendrán mediante el método de máxima verosimilitud. Estos estimadores serán más precisos, ya que se basan en una cantidad menor de parámetros, serán consistentes y asintóticamente más eficientes. El problema que surge,¡ es el error en escoger el modelo, ya que si se escoge incorrectamente, las estimaciones convergerán en valores equivocados.

El modelo de vida acelerada se establece cuando hay un modelo paramétrico subyacente:

$$\boldsymbol{S(t|Z) = S_0(t \exp(\theta' Z))} \text{ para todo } \boldsymbol{t > 0}$$

Siendo $\boldsymbol{T}$ el tiempo hasta $\boldsymbol{\varepsilon}$, $\boldsymbol{Z' = (Z_1, ..., Z_p)}$ un vector de covariables que no varían en el tiempo, $\boldsymbol{\theta' = (\theta_1, (...), \theta_p)}$ un vector de coeficientes de regresión y $\boldsymbol{S_0(t)}$ la supervivencia subyacente que corresponde a un individuo con $\boldsymbol{Z = 0}$.

Para medir el cambio de escala de tiempo producido por las covariables se calculará el **factor de aceleración** ($\boldsymbol{\exp(\theta' Z)}$). Este valor nos ayudará a ver que factores tienen un mayor riesgo de muerte cuando se interpreten los resultados del modelo.

3.3.1. Expresión log-lineal de modelo de Vida Acelerada

El modelo log-lineal, establece una relación lineal entre el logaritmo del tiempo y los valores de las covariables. Siendo $\boldsymbol{W}$ la distribución del error y dando valores reales superiores a $\boldsymbol{\mu}$, $\boldsymbol{\sigma > 0}$ y $\boldsymbol{\gamma' = (\gamma_1, (...), \gamma_p)}$, el modelo log-lineal para $\boldsymbol{T}$ tiene la siguiente relación:

$$\boldsymbol{Y = \ln T = \mu + \gamma' Z + \sigma W} \tag{3.1}$$

Este modelo sería equivalente al de vida acelerada con parámetro $\boldsymbol{\theta = -\gamma}$, para cualquier distribución $\boldsymbol{W}$.

3.3.2. Estimación de los parámetros por máxima verosimilitud

En un hipotético caso sin covariables, si observamos $\boldsymbol{(Y_1, \delta_1), (...), (Y_n, \delta_n)}$ donde $\boldsymbol{Y_i = \text{mín}\{\ln T_i, \ln C_i\}}$ y

$$\boldsymbol{\delta_i} = \begin{cases} \boldsymbol{1} & \text{si} \quad \boldsymbol{Y_i = \ln T_i} \quad \text{observación } \boldsymbol{i} \text{ no censurada} \\ \boldsymbol{0} & \text{si} \quad \boldsymbol{Y_i = \ln C_i} \quad \text{observación } \boldsymbol{i} \text{ censurada} \end{cases}$$

La función de máxima verosimilitud, cuando los datos sean censurados por la derecha, será la siguiente:

$$L(\mu,\sigma) = \prod_{j\in D} f_Y(y_j) \prod_{j\in R} S_Y(C_j)$$

$$= \prod_{i=1}^{n} \left\{ \left[\frac{1}{\sigma} f_W(\frac{y_1-\mu}{\sigma})\right]^{\delta_i} \left[S_W(\frac{y_1-\mu}{\sigma})\right]^{1-\delta_i} \right\}$$

$$= \prod_{i=1}^{n} \left\{ \left[\frac{1}{\sigma}\exp\left[\frac{y_1-\mu}{\sigma} - e^{\frac{y_1-\mu}{\sigma}}\right]\right]^{\delta_i} \left[\exp\left[-e^{\frac{y_1-\mu}{\sigma}}\right]\right]^{1-\delta_i} \right\}$$

Los estimadores de máxima verosimilitud para μ y σ, se denotarán como $\hat{\mu}$ y $\hat{\sigma}$ y son los valores que maximizan la función L(μ, σ).

3.3.3. Distribución de Weibull

En este trabajo se probará el uso de la distribución de Weibull para ajustar un modelo de supervivencia paramétrico, esta distribución es la única que admite una representación de riesgos proporcionales y de vida acelerada con el factor de aceleración $e^{-\gamma' Z}$.

Antecedentes

La primera propuesta de la distribución de Weibull fue dada por Rosen y Rammler en 1933, 6 años más tarde, Waloddi Weibull discute dicha propuesta en distintas situaciones de fallo. Esta distribución ha sido comúnmente usada para modelar la mortalidad específica de una mortalidad.

Gumbel en su libro *Statistics of Extremes*, conoce esta distribución como la primera distribución asintótica de los valores extremos (ver GUMBEL [11]).

Ley de Weibull

La distribución de Weibull con parámetro de escala p e índice k, proviene de la distribución exponencial cuando la función de riesgo depende de una potencia del tiempo:

$$\lambda(t) = kp(pt)^{k-1} \text{ para todo } t > 0$$

En caso que k sea igual a 1, la distribución de Weibull es equivalente a la distribución exponencial con función de riesgo: $\lambda(t) = p$.

La expresión de la función de supervivencia equivaldrá a:

$$S(t) = \exp\left[-p(t)^k\right]$$

La expresión de la función de densidad vendría dada por:

$$f(t) = kp(tp)^{k-1} \exp\left[-(pt)^k\right]$$

La expresión de la función de riesgo acumulado será igual a:

$$\Lambda(t) = p(t)^k$$

En caso que se suponga que el tiempo de supervivencia se ajusta a una distribución Weibull, el término del error en la expresión log-lineal (ver ecuación (3.1) en la página 37) sigue una distribución Gumbel estándar.

3.3.4. Distribución Log-normal

En este trabajo también se probará el uso de la distribución log-normal para ajustar el modelo paramétrico con nuestros datos. Esta distribución ha podido usarse con éxito en varios estudios de cáncer gracias a su pronunciada asimetría hacia la derecha, por eso, se comprobará si es útil para nuestros datos.

Un ejemplo de estudio seria el siguiente: **Log-normal**. En este artículo se explica como se ha medido la supervivencia de pacientes de cáncer de colon, a través de Kaplan-Meier, modelo de Cox y modelo log-normal. Se ha obtenido suficiente evidencia en los datos para decir que el modelo log-normal hace una mejor estimación de la supervivencia que los otros. En el artículo **log-normal2** se prueba la eficiencia del modelo log-normal para pacientes de cáncer (en este caso de pulmón), este modelo, muestra una eficiencia mayor al de Cox y Weibull.

Para una ley log-normal de parámetros p y τ la función de densidad se expresará como:

$$f(t) = \frac{1}{\tau t\sqrt{(2\pi)}} \exp\left(-\frac{[\ln(tp)]^2}{2\tau^2}\right)$$

La función de supervivencia será escrita como:

$$S(t) = 1 - \Psi\left(\frac{1}{\tau}\ln pt\right)$$

Y la función de riesgo se escribe como:

$$\lambda(t) = \frac{\frac{1}{\tau t\sqrt{(2\pi)}} \exp\left(-\frac{[\ln(tp)]^2}{2\tau^2}\right)}{1 - \Psi\left(\frac{\ln pt}{\tau}\right)}$$

Si se supone que el tiempo de supervivencia sigue una distribución log-normal, el término del error en la expresión log-lineal (ver ecuación (3.1) en la página 37) sigue una distribución normal estándar.

3.4. Modelo de riesgos proporcionales de Cox

El uso del modelo de Cox o modelo de riesgos proporcionales se introdujo en 1972 por David Roxbee Cox. Este modelo es el más utilizado en el ámbito del análisis de supervivencia cuando hay datos censurados. La principal ventaja de este modelo respecto a los demás es que no se necesita hacer ninguna suposición sobre la distribución de $\boldsymbol{T}$. En el ámbito de supervivencia es un equivalente a una regresión lineal, el parecido que tiene con una regresión logística se encuentra en las tasas de riesgo.

3.4.1. Expresión del modelo

La principal asunción del modelo de riesgos proporcionales es que es constante durante un período de tiempo y el efecto de las covariables es relacionado linealmente con el logaritmo de la razón de riesgos.

La función de riesgo para el modelo de Cox tiene la siguiente expresión:

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

Donde $\boldsymbol{\lambda_0}$ representa la función de riesgo basal, es decir cuando todas las covariables son iguales a 0. Por otro lado, $\boldsymbol{X_1, ..., X_p}$ son covariables fijas. El conjunto de todas las covariables se denotará como $\boldsymbol{X = (X_1, X_2, ..., X_p)'}$.

En este modelo se supone que la razón entre las funciones de riesgo se mantiene contante a lo largo del tiempo, por lo tanto, se verifica:

$$\boldsymbol{\frac{\lambda(t|Z)}{\lambda_0(t)} = \exp\{\beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_p Z_p\}}$$

Se puede comprobar cómo el término de la derecha sólo depende de los valores de las covariables y no del tiempo $\boldsymbol{t}$. Por otro lado, la distribución del error no ha sido especificada.

El factor $\boldsymbol{\exp\{\beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_p Z_p\}}$ corresponderá al *hazard ratio* de un individuo $\boldsymbol{Z}$ respecto a un individuo $\boldsymbol{Z = 0}$ y expresa cuántas veces es mayor el riesgo instantáneo de padecer el evento de interés con un perfil $\boldsymbol{Z}$ en comparación a un perfil $\boldsymbol{Z = 0}$. Este valor se usará para comprobar qué factores de supervivencia han sido los más determinantes.

3.4.2. Función de verosimilitud parcial

En el modelo de riesgos proporcionales no es posible la estimación de los parámetros mediante la función de verosimilitud, ya que no se hace ninguna suposición de la función de riesgo parcial; por este motivo, se utiliza la función de verosimilitud parcial.

La función de verosimilitud parcial es definida como:

$$\boldsymbol{L(\beta_1, ..., \beta_p) = \prod_{j=1}^{r} \text{Prob}\{e_j = i|\Gamma_j\} = \prod_{j=1}^{r} \text{Prob}\{Z_{(j)} = z_{(j)}|\Gamma_j\}}$$

Donde $\boldsymbol{r}$ número de eventos $\boldsymbol{\varepsilon}$ (muerte a causa de cáncer colorrectal). $\boldsymbol{t_1, ..., t_r}$ serán los tiempos de muerte ordenados. $\boldsymbol{\Gamma = (Y_i, \delta_i, X_i)}$ corresponderá al conjunto de información en la muestra, $\boldsymbol{\Gamma_j}$ se entenderá como el conjunto de información hasta el momento $\boldsymbol{t_{(j)}}$. $\boldsymbol{Z_{(j)}}$ será la variable aleatoria que corresponderá con los valores del vector de covariables para el individuo muerto en $\boldsymbol{t_{(j)}}$. $\boldsymbol{z_{(j)}}$ corresponderá al vector de covariables. Por último, $\boldsymbol{e_i}$ indica a que individuo corresponde la muerte.

La maximización de esta función permitirá la estimación de los coeficientes $\boldsymbol{\beta_j}$. Generalmente estos estimadores cumplirán las propiedades del método de máxima verosimilitud. La gran controversia que causó en sus inicios esta función fue debida a que no depende unicamente de los $\boldsymbol{\beta_j}$, sino que también de la función $\boldsymbol{\lambda_0(j)}$, la cual no está parametrizada.

3.4.3. Validación del modelo

En el modelo de riesgos proporcionales, la validez del modelo y el análisis de los residuos tiene una gran importancia. Esta validez está en gran parte medida por la validación de los riesgos proporcionales. Esto se puede comprobar de forma analítica o gráfica mediante los residuos.

Los residuos se calculan para cada paciente y proporcionan información sobre el valor observado y el estimado por la ecuación de regresión obtenida en el modelo.

Residuos

Hay diferentes definiciones de residuos, existen los residuos de ***Cox y Snell***, los basados en la ***deviance***, los residuos basados en ***scores***, los que están basados en ***martingalas*** y los residuos de ***Schoenfeld***.

A continuación se explicarán los residuos de ***Schoenfeld*** y los basados en ***martingalas***, los primeros se usarán para validar el modelo y la hipótesis de riesgos proporcionales, los segundos serán usados para decidir cuál es la mejor forma funcional para una covariable concreta.

Residuos de Schoenfeld

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

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

Para esta expresión $\boldsymbol{J_i(t)}$ es el indicador de que el individuo $\boldsymbol{j}$ permanece en el estudio en el momento justo antes del tiempo $\boldsymbol{t}$, $\boldsymbol{\delta_i}$ será el indicador de censura, $\boldsymbol{X_{ik}}$ es el valor de la k-ésima covariable del individuo $\boldsymbol{i}$ y $\boldsymbol{\bar{X}_k(T_i)\}}$ es el valor que corresponde al promedio de la covariable $\boldsymbol{k}$ en el tiempo $\boldsymbol{T_i}$.

Comprobación gráfica de la proporcionalidad de riesgos

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

Para comprobar gráficamente que se cumple esta premisa para una covariable determinada, los residuos se deberán agrupar de forma aleatoria a ambos lados del valor 0 del eje Y, y la curva del estimador $\boldsymbol{\beta}$ de esa covariable debe ser de pendiente 0, una línea recta.

Como ejemplo, se insertará el Gráfico 3.1 donde se puede deducir del gráfico que la covariable cumple la premisa, ya que la estimación de beta en función del tiempo se mantiene casi constante.

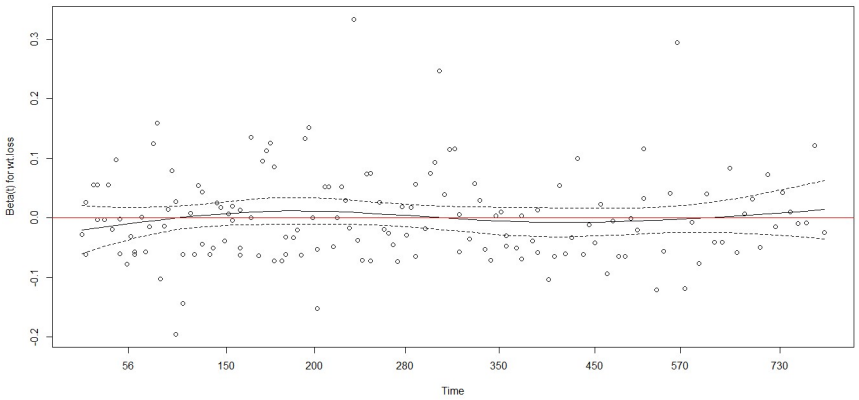


Figura 3.1: Ejemplo de los residuos de Schoenfeld para la variable *wt.loss* de la base de datos *lungs* del paquete *survival* en R.

Residuos basados en martingalas

Los residuos basados en ***martingalas*** son definidos para cada individuo $\boldsymbol{i}$ de la siguiente forma:

$$\boldsymbol{r_{M_i} = \delta_i\text{-}r_{C_i} = 0} \begin{cases} \mathbf{1 - r_{C_i} = 1 - \exp\{\hat{\beta}' Z_i\}\hat{\Lambda}_0(y_i)} & \text{si la observación no está censurada} \\ \mathbf{-r_{C_i} = -\exp\{\hat{\beta}' Z_i\}\hat{\Lambda}_0(y_i)} & \text{si la observación está censurada} \end{cases}$$

Estos residuos se pueden interpretar como diferencias entre el número de eventos observados y el valor del esperado bajo el modelo de Cox.

Comprobación gráfica

Estos residuos se usarán para comprobar gráficamente cual es la mejor transformación de una covariable de forma que esta explique de manera óptima su efecto sobre la supervivencia. Cuando el gráfico de los residuos de martingalas resulte lineal, no hará falta ninguna transformación, en caso contrario habría que estudiar

cuál es la mejor transformación de la covariable en cuestión.

En el Gráfico 3.2 se pueden ver los residuos de una covariable que tiene una transformación óptima en el modelo de Cox correspondiente.

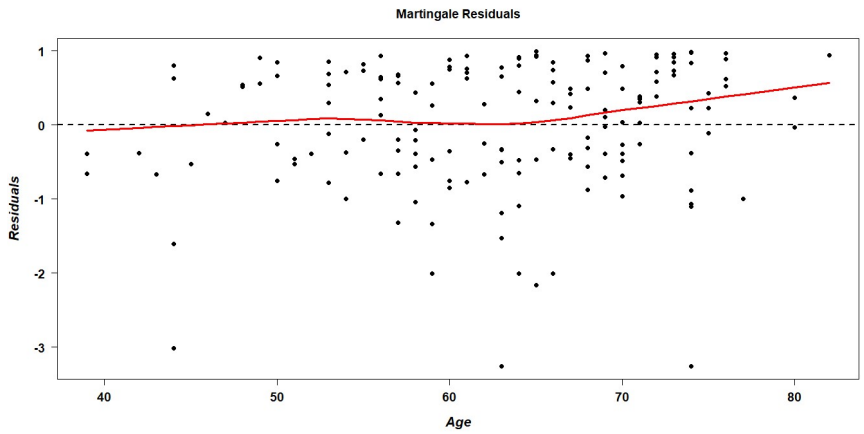


Figura 3.2: Ejemplo de los residuos basados en Martingala para la variable *age* de la base de datos *lungs* del paquete *survival* en R.

Capítulo IV

Resultados

En este capítulo se procederá a analizar la supervivencia de las dos bases de datos, este análisis se hará sobre una tasa de supervivencia hasta los 5 años o 60 meses, para ello, se ha creado una nueva variable de supervivencia y otra de censura a partir de las dos variables que había en la base de datos, el código usado se puede ver en el Código 3.

```
dades1$survival5ys <- pmin(dades1$SURVIV_MONTHS, 5 * 12)
dades1$censura5ys <- ifelse(dades1$SURVIV_MONTHS > 5 * 12, 0, dades1$SEER_DEATH_
    CAUSE == 1)

dades2$survival5ys <- pmin(dades2$SURVIV_MONTHS, 5 * 12)
dades2$censura5ys <- ifelse(dades2$SURVIV_MONTHS > 5 * 12, 0, dades2$SEER_DEATH_
    CAUSE == 1)
```

Código 3: Creación de nuevas variables hasta los 5 años en R.

El siguiente análisis se dividirá en tres partes, la primera corresponderá al análisis no paramétrico mediante el uso de Kaplan y Meier. En este apartado, se verá qué influencia tiene cada variable categórica en la curva de supervivencia y la forma de dicha curva. Se han hecho más análisis que no se incluirán, en este apartado se comentará la información que se ha considerado más relevante.

En la segunda y tercera parte se presentarán distintos modelos formados con las variables **Género**, **Etnia**, **Grado**, **Estado civil** y **Edad** para la base de datos del período temporal entre 1975 y 1984, para la base de datos correspondiente al intervalo de años entre 2004 y 2011 se usarán las mismas variables y se añadirán variables de **Seguro Médico** y **Tamaño del tumor**. Se eliminarán los datos de todas las categorías que sean igual a *Desconocido*, ya que no aportan ninguna información y no se ha encontrado ninguna similitud con ninguna otra categoría. En la variable **Seguro Médico** no se ha eliminado la categoría *Desconocido*, porque se perdería mucha información.

Para la segunda parte de este capítulo se observará qué distribución (log-normal o Weibull) se ajusta mejor a la supervivencia de los datos para cada base de datos. Una vez decidido, se construirá un modelo de regresión

con la distribución paramétrica que mejor se ajuste, se mostrarán los resultados de los principales estadísticos y se realizará la validación del modelo.

Además, en la tercera parte se ajustará una regresión mediante el modelo de Cox, se mostrarán los resultados obtenidos y se validará la premisa de riesgos proporcionales para cada covariable.

El nivel de significación que se establecerá será de $\boldsymbol{\alpha = 0{,}05}$.

4.1. Análisis no paramétrico

Como se ha comentado anteriormente, se usará el método de Kaplan y Meier para este análisis. Para empezar, se presentarán las tablas con la información del número de personas, eventos, valor de la mediana y el valor del cuantil del 25 % para cada categoría. También se hará la prueba del Log-rank para comprobar la igualdad de supervivencia de las distintas categorías de la variable. Se mostrará la curva de supervivencia global para cada base de datos y distintas curvas de supervivencia para variables categóricas, después se comentarán los resultados más relevantes. Por último, se hará una breve comparación entre los resultados de las dos bases de datos.

Se decidió añadir el cuantil del 25 %, el cuál indica el tiempo que pasa hasta llegar a una probabilidad de supervivencia del 75 %, en el momento que se detectó que la gran mayoría de categorías tenían una supervivencia mayor al 50 % en los 60 meses del estudio.

4.1.1. Período 1975-1984

En la Tabla 4.1 se presentarán los resultados del estimador Kaplan y Meier para todas las variables categóricas de la base de datos de los pacientes diagnosticados entre 1975 y 1984.

Categorías	**n**	**Eventos**	**Mediana**	**Cuantil 25 %**	**Prueba Log-rank chisq (p-valor)**
	61316	25905	-	15	
Género					23.7 (1e-06)
Hombre	30786	13238	-	15	
Mujer	30530	12667	-	14	
Etnia					58.2 (1e-12)
Blanca	54758	23015	-	15	
Afroamericana	3984	1853	58	11	
Asiática o isleña	2415	980	-	21	
Otras	159	57	-	29	

Categorías	n	Eventos	Mediana	Cuantil 25 %	Prueba Log-rank chisq (p-valor)
Grado					2498 (0)
Grado I	13949	4605	-	29	
Grado II	33945	13764	-	20	
Grado III	12117	6697	30	8	
Grado IV	1305	839	15	3	

Tabla 4.1: Descriptiva no paramétrica de Supervivencia período 1975-1984.

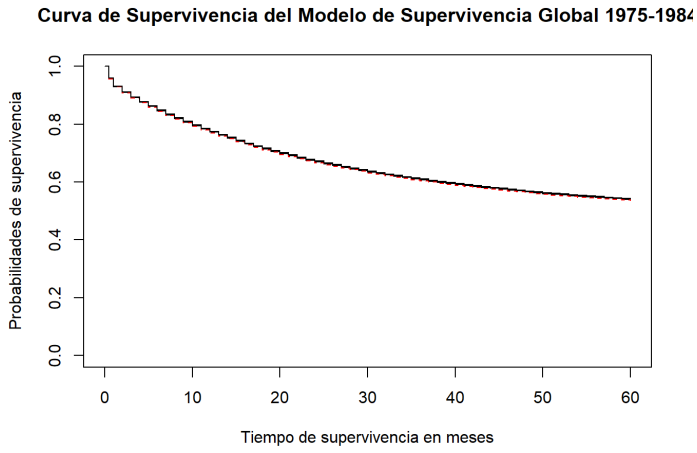
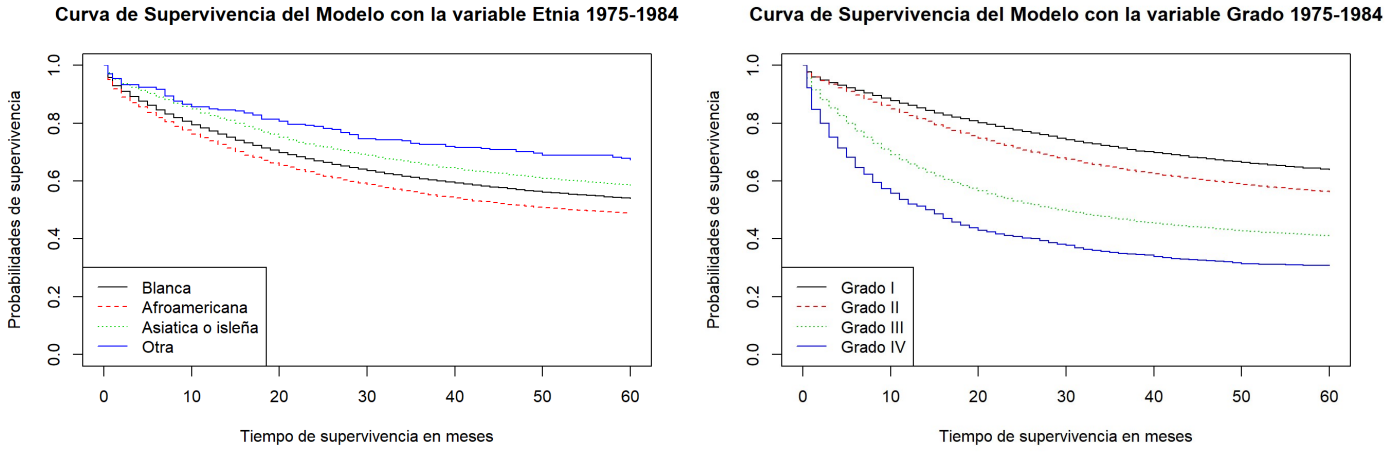


Figura 4.1: Curva de supervivencia del modelo de supervivencia global para la base de datos de 1975-1984.



(a) Grado

(b) Seguro médico

Figura 4.2: Curvas de supervivencia de los modelos con la variable Etnia y Grado para la base de datos 1975-1984

Cabe destacar en este primer análisis de supervivencia, que la gran mayoría de categorías y el modelo global (ver Gráfico 4.1) no llegan a una probabilidad de supervivencia menor al 50 %. Las únicas categorías que han llegado han sido *Grado III* y *Grado IV* de la variable **Grado** y la categoría *Afroamericana* de la variable **Etnia**.

En el Gráfico 4.2b se puede comprobar como hay una gran diferencia entre las personas que han padecido un cáncer de grado I y II o las que lo han sufrido de III o IV, las segundas tienen una probabilidad de supervivencia muy inferior. De hecho, las que han padecido un cáncer de grado IV tienen una probabilidad de supervivencia inferior al 75 % a partir de los 3 meses y menor al 50 % a partir de los 15 meses de estudio.

En el Gráfico 4.2a se observa como los afroamericanos tienen una supervivencia inferior a las otras etnias. A pesar de esto, los afroamericanos no bajan de una probabilidad de supervivencia del 50 % hasta los 58 meses, también se puede ver como en 11 meses ya tienen una probabilidad de supervivencia inferior al 75 %.

Se han realizado pruebas de logrank para todas las variables categóricas, en todas ellas se ha obtenido un p-valor inferior al nivel de significación, hay que tener en cuenta que debido a la gran catidad de datos, incluso diferencias pequeñas en la muestra pueden resultar estadísticamente significativas. Según estos p-valores hay suficiente evidencia para rechazar la hipótesis nula, por lo tanto, las categorías no son estadísticamente iguales entre sí.

4.1.2. Período 2004-2011

En la Tabla 4.2 se presentarán los resultados del estimador Kaplan y Meier para las variables categóricas de la base de datos del intervalo de tiempo correspondiente entre 2004 y 2011.

Categorías	**n**	**Eventos**	**Mediana**	**Cuantil 25 %**	**Prueba Log-rank chisq (p-valor)**
	64557	19023	-	32	
Género					0.8 (0.4)
Hombre	33023	31534	-	35	
Mujer	9729	9294	-	29	
Etnia					201 (0)
Blanca	50441	14554	-	32	
Afroamericana	6889	2502	-	24	
Asiática o isleña	6528	1780	-	43	
Otras	699	187	-	-	
Grado					2743 (0)
Grado I	6720	1123	-	-	
Grado II	44400	11907	-	46	
Grado III	11974	5332	-	13	
Grado IV	1463	661	-	11	

Categorías	n	Eventos	Mediana	Cuantil 25 %	Prueba Log-rank chisq (p-valor)
Seguro médico					211 (0)
No asegurado	895	325	-	18	
Seguro público	3606	1333	-	19	
Asegurado	34198	9554	-	36	
Desconocido	865	269	-	19	

Tabla 4.2: Descriptiva no paramétrica de Supervivencia período 2004-2011.

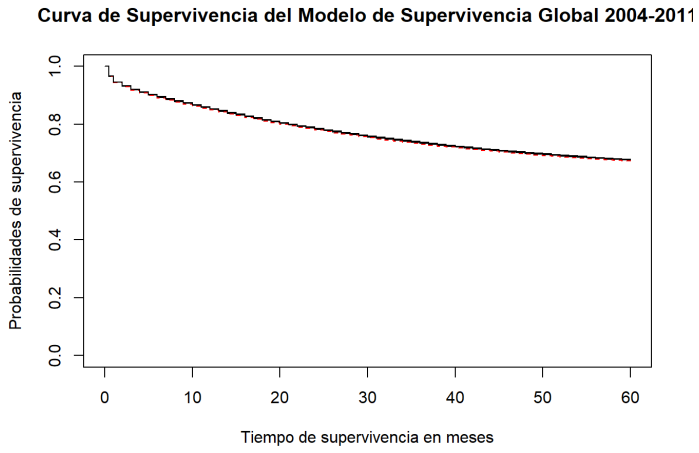
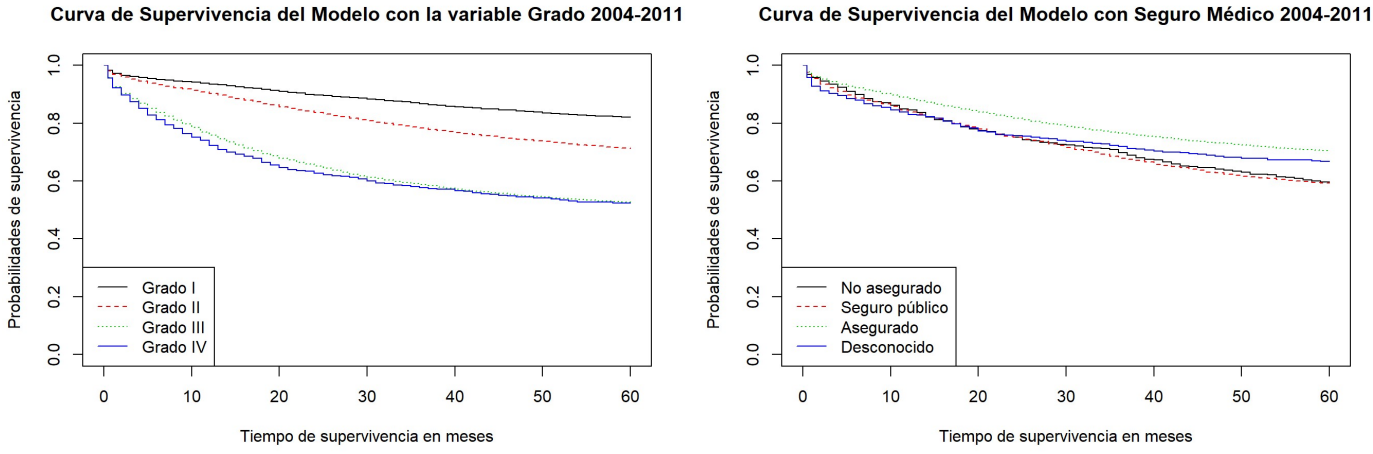


Figura 4.3: Curva de supervivencia del modelo de supervivencia global para la base de datos de 1975-1984.



(a) Grado

(b) Seguro médico

Figura 4.4: Curvas de supervivencia de los modelos con la variable **Grado** y **Seguro Médico** para la base de datos 2004-2011

El segundo análisis no paramétrico se ha realizado sobre la base de datos correspondiente al intervalo temporal entre los años 2004 y 2011. Todas las categorías tienen una probabilidad de supervivencia superior al 50 % durante los primeros 60 meses de estudio, se puede observar en el Gráfico 4.3.

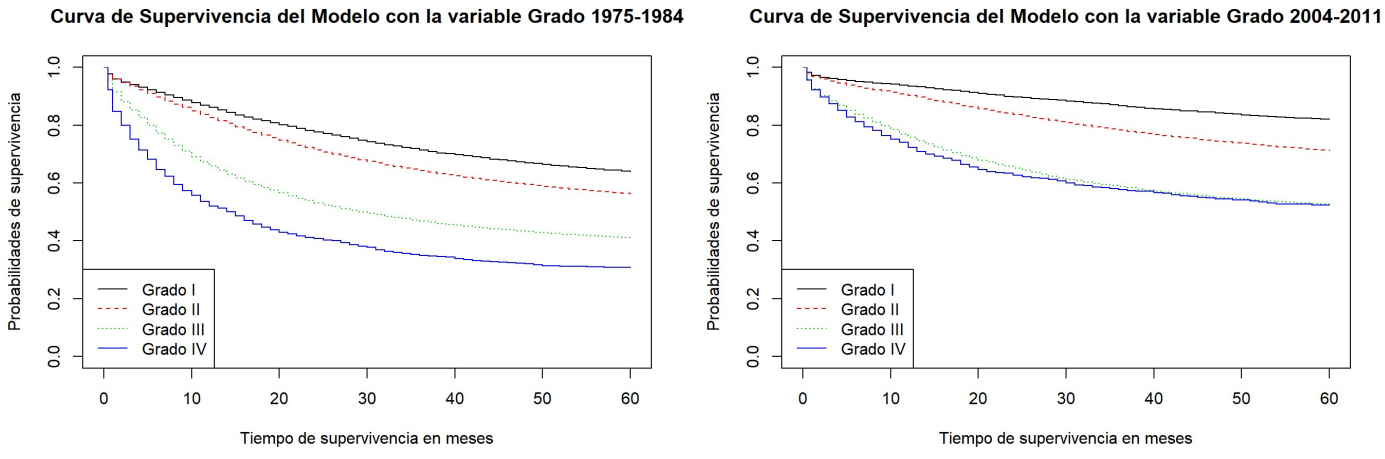
En el Gráfico 4.4a se puede observar como para la variable **Grado** como las categorías *Grado III* y *Grado IV* se distribuyen prácticamente igual y su probabilidad de supervivencia es bastante más baja que la de las categorías que corresponden a las personas que han padecido un cáncer de grado I y II.

En el Gráfico 4.4b para la variable **Seguro Médico**, las curvas de supervivencia de las categorías *Seguro Público* y *No asegurado* son prácticamente iguales. Además, la diferencia entre estas dos categorías y la *Asegurado* se va incrementando a lo largo del tiempo. Por lo tanto, la diferencia entre la probabilidad de supervivencia entre estas categorías es mucho más alta al cabo de 50 meses del inicio del estudio que en los primeros 10 meses.

Las pruebas de log-rank realizadas han dado como resultado un p-valor inferior al nivel de significación excepto para la variable género. Según estos p-valores hay suficiente evidencia para rechazar la hipótesis nula, por lo tanto, las categorías no son estadísticamente iguales entre sí. Para la variable género no hay suficiente evidencia para rechazar la hipótesis nula.

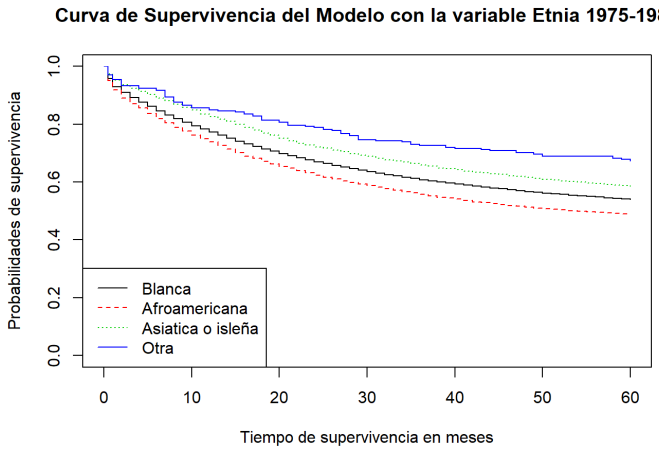
4.1.3. Comparación de ambos períodos

En este apartado, se comentarán las diferencias observadas entre las curvas de supervivencia de las dos bases de datos para las variables que se han considerado más relevantes.

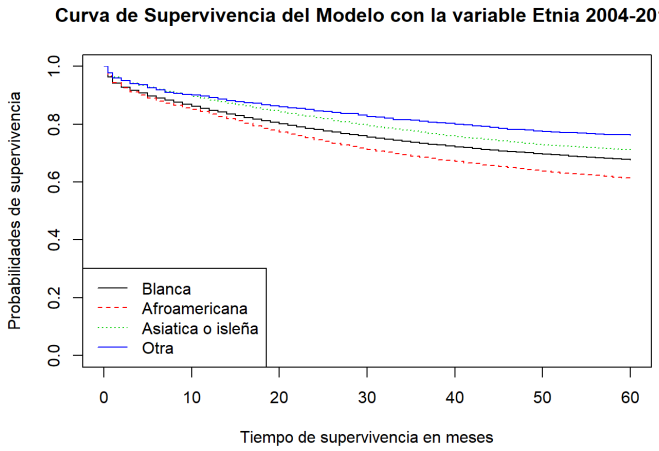


(a) Grado 1975-1984 **(b)** Grado 2004-2011

Figura 4.5: Curvas de supervivencia de los modelos con la variable **Grado** para las bases de datos 1975-1984 y 2004-2011



(a) Etnia 1975-1984



(b) Etnia 2004-2011

Figura 4.6: Curvas de supervivencia de los modelos con la variable **Etnia** para las bases de datos 1975-1984 y 2004-2011

Observando los resultados obtenidos en las Tablas 4.1 (página 46) y 4.2 (página 48), se puede ver cómo en la base de datos de pacientes diagnosticados entre 2004 y 2011, todas las categorías sobreviven un mayor número de meses hasta bajar del 75 % de probabilidad de supervivencia.

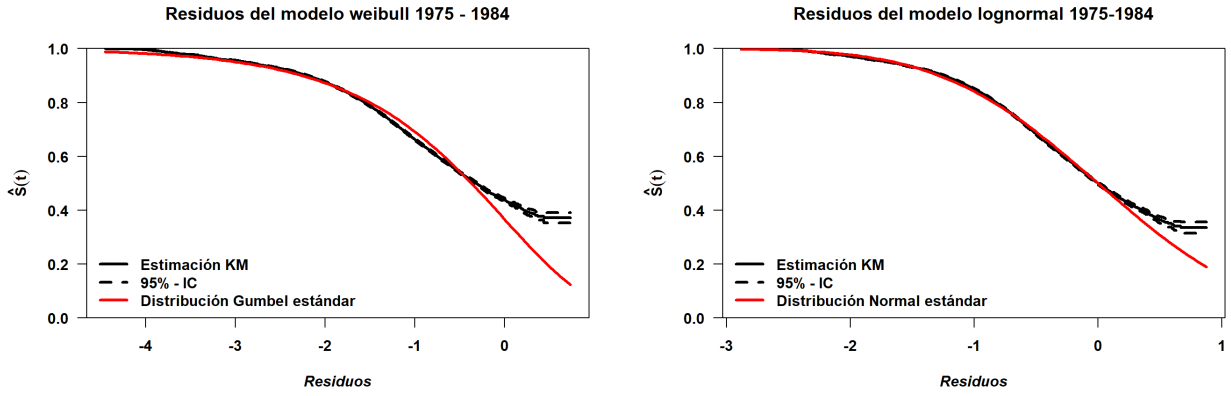
Se puede observar en los Gráficos 4.5a y 4.5b co¡ómo en la variable **Grado** ha habido una gran mejora de la probabilidad de supervivencia para todas las categorías, también se ha visto como la probabilidad de supervivencia entre las categorías de *Grado III* y *Grado IV* ha pasado de tener una visible diferencia a ser prácticamente igual, teniendo en cuenta que para ambas categorías esta probabilidad ha mejorado. Además, la diferencia de las probabilidades de supervivencia entre las categorías *Grado I* y *Grado II* con las categorías *Grado III* y *Grado IV* ha disminuido.

En los Gráficos 4.6a y 4.6b se observa como para la variable Etnia, todas las categorías han mejorado su probabilidad de supervivencia en la base de datos correspondiente al intervalo de tiempo entre el año 2004 y 2011, además parece que las diferencias entre ellas se mantienen de forma muy similar. Sin embargo, si se presta atención a las probabilidades de supervivencia correspondientes a los tiempos de supervivencia posteriores a los 30 meses del inicio del estudio, las diferencias de dichas probabilidades aumentan de forma gradual entre las personas de etnia afroamericana y las demás. Este hecho no se observa en la base de datos 1975-1984.

4.2. Modelo de vida acelerada

4.2.1. Período 1975-1984

Los siguientes gráficos se han realizado ajustando un modelo de supervivencia K-M con los residuos del modelo de regresión y la variable de censura. Para el modelo de Weibull se ha añadido una curva de la distribución de Gumbel estándar, para el log-normal se ha añadido una curva de la distribución normal estándar.



(a) Weibull **(b)** Log-normal

Figura 4.7: Gráficas de los residuos para comprobar el mejor ajuste a la distribución paramétrica para la base de datos 1975-1984

Se puede comprobar en los Gráficos 4.7a y 4.7b como el modelo ajustado por una distribución log-normal es el que mejor se ajusta a los datos de la base de datos que corresponde al intervalo de tiempo entre 1975 y 1984. Por lo tanto, se analizarán los resultados correspondientes a una regresión ajustada por la distribución log-normal.

Variable	Categoría	Estimación del parámetro	Factor de aceleración	Error estándar	P-valor
Intercept		6.105	0.002	0.064	$< \mathbf{2e^{-16}}$
Género	Mujer	0.148	0.862	0.02	$\mathbf{3{,}8e^{-13}}$
Etnia	Afroamericana	-0.42	1.522	0.041	$< \mathbf{2e^{-16}}$
	Asiática o isleña	0.128	0.88	0.053	**0,015**
	Otras	0.3	0.741	0.205	**0,146**

Variable	Categoría	Estimación del parámetro	Factor de aceleración	Error estándar	P-valor
Grado	Grado II	-0.363	1.438	0.026	$< \mathbf{2e^{-16}}$
	Grado III	-1.315	3.725	0.031	$< \mathbf{2e^{-16}}$
	Grado IV	-2.019	7.531	0.068	$< \mathbf{2e^{-16}}$
Edad		-0.02	1.02	0.0009	$< \mathbf{2e^{-16}}$

Tabla 4.3: Ajuste del modelo de vida acelerada para la distribución log-normal para el período 1975-1984.

Los p-valores que se han obtenido en la Tabla 4.3, indican si hay suficiente evidencia para rechazar $\boldsymbol{H_0 : \gamma_j = 0}$. Esta hipótesis indica si el coeficiente obtenido por si solo tiene un efecto en el modelo o no. Por lo tanto, cuando los p-valores sean inferiores al nivel de significación habrá una evidencia estadística para decir que ese coeficiente tiene efecto en el modelo.

Primeramente, se tendrá en cuenta que las categorías de referencia para las distintas variables categóricas del modelo son *Hombre*, *Blanca* y *Grado I*. Además, es importante decir que para la interpretación de los resultados obtenidos se usará el valor del factor de aceleración.

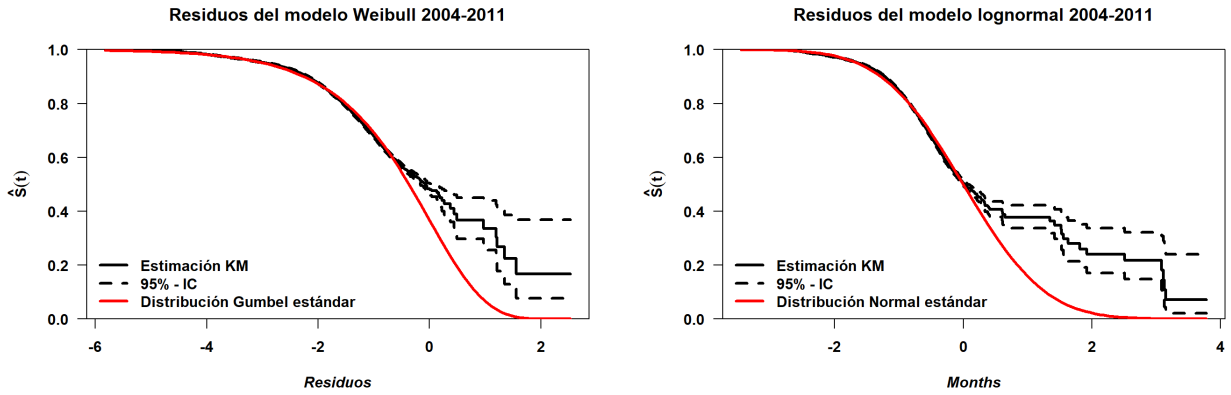
Según los resultados obtenidos en esta regresión se puede comprobar como las personas blancas de este estudio tienen una mediana de supervivencia 1.522 veces superior que las personas afroamericanas. Además, si se observa el p-valor obtenido este es inferior al nivel de significación, esto indica que hay suficiente evidencia para rechazar que esta diferencia no se cumple durante el tiempo.

Se observa como para la variable **Grado** igual que se ha visto en las curvas de supervivencia ajustadas por Kaplan y Meier, los pacientes que han padecido un cáncer de grado superior tienen un menor tiempo de supervivencia. Por ejemplo, una persona que ha padecido un cáncer de grado I tendrá una mediana de supervivencia 7.5 veces superior a una persona que lo ha padecido de grado IV.

Si se presta atención en el factor de aceleración obtenido en la variable numérica **Edad**, se puede ver como una persona 20 años inferior a otra tendrá una mediana de supervivencia $\boldsymbol{e^{-\gamma * 20} = e^{0{,}02 * 20} = 1{,}523}$ veces superior.

4.2.2. Período 2004-2011

A continuación, se comprobará el ajuste que tienen los datos de los pacientes diagnosticados entre 2004 y 2011 para una distribución Weibull y para una log-normal.



(a) Weibull **(b)** Log-normal

Figura 4.8: Gráficas de los residuos para comprobar el mejor ajuste a la distribución paramétrica para la base de datos 2004-2011

En los Gráficos 4.8a y 4.8b, se observa como ninguna de las dos distribuciones se ajusta del todo bien para los residuos positivos más altos. A pesar de esto, se observa como la distribución de Weibull se ajusta mejor a esta base de datos (4.8a). Se analizarán los resultados del análisis de la regresión ajustada por la distribución de Weibull.

Variable	Categoría	Estimación del parámetro	Factor de aceleración	Error estándar	P-valor
Intercept		7.504	0.0006	0.123	$< \mathbf{2e^{-16}}$
Género	Mujer	0.108	0.898	0.029	**0,00022**
Etnia	Afroamericana	-0.458	1.581	0.044	$< \mathbf{2e^{-16}}$
	Asiática o isleña	0.081	0.922	0.049	**0,09669**
	Otras	0.064	0.938	0.145	**0,6596**
Grado	Grado II	-0.6	1.822	0.064	$< \mathbf{2e^{-16}}$
	Grado III	-1.54	4.665	0.068	$< \mathbf{2e^{-16}}$
	Grado IV	-1.735	5.669	0.094	$< \mathbf{2e^{-16}}$
Seguro Médico	Seguro público	0.088	0.916	0.097	**0,364**
	Asegurado	0.635	0.53	0.089	$\mathbf{1{,}2e^{-12}}$
	Desconocido	0.499	0.607	0.136	**0,00023**

Variable	Categoría	Estimación del parámetro	Factor de aceleración	Error estándar	P-valor
Tamaño del tumor		-0.005	1.005	0.0001	$< \mathbf{2e^{-16}}$
Edad		-0.023	1.023	0.001	$< \mathbf{2e^{-16}}$

Tabla 4.4: Ajuste del modelo de vida acelerada para la distribución de Weibull para el período 2004-2011.

Para la regresión de esta base de datos, las categorías de referencia serán *Hombre*, *Blanca*, *Grado I* y *No asegurado*.

Si se presta atención al factor de aceleración de la variable género de la Tabla 4.4, se puede ver como las mujeres tienen una mediana de supervivencia $\frac{1}{\mathbf{0,898}} = \mathbf{1,114}$ veces más alta que los hombres.

Para la variable **Seguro Médico** cabe destacar, que las personas que han participado en este estudio y están aseguradas, tienen una mediana de supervivencia $\frac{1}{\mathbf{0,53}} = \mathbf{1,89}$ veces más alta que las personas que no gozaban de un seguro médico.

Observando el factor de aceleración de la variable **Tamaño del Tumor**, un paciente que tenga un tumor 30 milímetros inferior a otro tendrá una mediana de supervivencia $\boldsymbol{e^{0,005*30} = 1,16}$ veces superior.

4.2.3. Comparación entre ambos períodos

En este apartado se analizará las diferencias de resultados que ha habido entre los modelos de vida acelerada realizados para las distintas bases de datos.

El ajuste de la base de datos del 1975-1984 es mejor a la distribución paramétrica log-normal (Gráfico 4.7b, página 51) que el ajuste de los datos de la base de datos de pacientes diagnosticados entre 2004 y 2011 para la distribución Weibull (Gráfico 4.8a, página 53).

Observando los factores de aceleración de la variable **Etnia**, se puede ver como en la base de datos más reciente a la actualidad (Tabla 4.4, página 54), la diferencia entre tiempos de supervivencia entre las categorías *Asiática o isleña* y *Otras* con la categoría de referencia blanca ha disminuido. En la primera base de datos una persona asiática o isleña tenía una mediana de supervivencia $\frac{1}{\mathbf{0,88}} = \mathbf{1,14}$ veces superior a las blancas (Tabla 4.3, página 52), mientras que en la segunda base de datos es $\frac{1}{\mathbf{0,92}} = \mathbf{1,09}$ veces superior (Tabla 4.4, página 54). Esto contrasta con la diferencia de tiempos de supervivencia entre las categorías referentes a personas afroamericanas y blancas, estas en vez de disminuir han aumentado. En la primera base de datos la mediana de supervivencia de personas blancas era 1.522 veces superior y en la base de datos más actual es 1.581 veces. Para poder afirmar dicha diferencia con más seguridad, se tendría que ajustar un único modelo que incluyera la variable período (1975-1984 vs. 2004-2011) y la interacción de esta variable con las demás.

En la variable **Grado** de la base de datos actual (Tabla 4.4, página 54) se observa la mejoría que ya se comentó en el apartado anterior respecto a los tumores de *Grado IV*. Este factor de aceleración ha bajado considerablemente respecto al de referencia en comparación al que hay en la base de datos correspondiente

al intervalo de tiempo entre 1975 y 1984 (Tabla 4.3, página 52). En la base de datos más actual una persona que padece el cáncer de grado IV tiene una mediana de supervivencia 5.669 veces inferior a una persona que lo padece de grado I, en la base de datos anterior es 7.53 veces. Es destacable como para los tumores de grado II y sobretodo de grado III esta diferencia ha crecido en la base de datos más reciente. Por ejemplo, una persona que ha participado en el estudio entre 2004 y 2011 con un tumor de grado I tiene una mediana de supervivencia 4.665 veces superior a una persona con un tumor de grado III, en cambio, una persona con ese mismo tumor que participó en el estudio entre 1975 y 1984 tenia una mediana de supervivencia 3.725 veces inferior a una persona con un tumor de grado I.

4.3. Modelo de Cox

A continuación, se mostrarán los resultados del análisis de supervivencia según el modelo de riesgos proporcionales de Cox para cada una de las bases de datos. Como se ha comentado en el capítulo anterior, la expresión del modelo es la siguiente:

$$\boldsymbol{\lambda(t|\mathrm{X}) = \lambda_0(t)\exp\{\beta'\mathrm{X}\} = \lambda_0(t)\exp\{\beta_1\mathrm{X}_1 + ... + \beta_i\mathrm{X}_i\}}$$

4.3.1. Período 1975-1984

En primer lugar, se comprobará a través de los residuos de martingalas en el Gráfico 4.9 si incluir de forma lineal la variable **Edad** en el modelo es la mejor opción para nuestros datos.

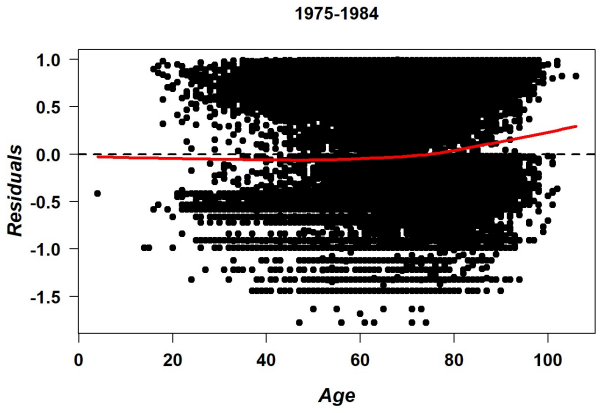


Figura 4.9: Gráfica de los residuos de Martingalas para la variable **Edad** para la base de datos 1975-1984

Se observa en el Gráfico 4.9 como la línea roja es horizontal en el eje 0. Hay una tendencia ascendente a partir de los 80 años, pero no parece indicar que incluir la edad de forma lineal en el modelo sea incorrecto.

Al incluirse la variable **Edad** sin categorizar, en la base de datos de los pacientes que fueron diagnosticados de cáncer colorrectal entre los años 1975 y 1984, se incluirán 7 covariables en el modelo. Las categorías de

referencia serán igual que en el modelo de vida acelerada (*Hombre*, *Blanco* y *Grado I*).

La expresión del modelo quedaría así:

$$\boldsymbol{\lambda(t|\mathrm{X}) = \lambda_0(t)\exp\left\{\sum_{i}^{7}\beta_i \mathrm{X}_i\right\}}$$

Variable	Categoría	Coeficiente	HR	0.95 IC	P-valor
Género	Mujer	-0.102	0.903	**[0,881 , 0,925]**	$\mathbf{2{,}82e^{-16}}$
Etnia	Afroamericana	0.248	1.282	**[1,222 , 1,345]**	$\mathbf{< 2e^{-16}}$
	Asiática o isleña	-0.053	0.948	**[0,889 , 1,011]**	**0,1045**
	Otras	-0.233	0.792	**[0,881 , 0,925]**	**0,0785**
Grado	Grado II	0.254	1.289	**[1,247 , 1,333]**	$\mathbf{< 2e^{-16}}$
	Grado III	0.781	2.183	**[2,102 , 2,267]**	$\mathbf{< 2e^{-16}}$
	Grado IV	1.164	3.205	**[2,978 , 3,451]**	$\mathbf{< 2e^{-16}}$
Edad		0.009	1.009	**[1,008 , 1,01]**	$\mathbf{< 2e^{-16}}$

Tabla 4.5: Ajuste del modelo de Cox para el período 1975-1984.

En este modelo ha habido un total de 25905 eventos y se han incluido 61316 pacientes.

Como se puede ver en la Tabla 4.5, todas las variables son estadísticamente significativas según el nivel de significación que se ha fijado excepto las categorías *Asiática o isleña* y *Otras*, esto probablemente sea debido al inferior número de datos que se tienen de estas dos categorías, también se puede observar como el parámetro de dichas covariables es muy cercano a 0.

Si observamos el riesgo instantáneo (HR) para la variable **Etnia**, se ven unos resultados similares a los de los análisis anteriores, las personas afroamericanas son las que tienen un mayor riesgo de fallecimiento por cáncer colorrectal. Por ejemplo, una persona afroamericana tiene un riesgo instantáneo de morir 1.282 veces más alto que una persona blanca. Por otro lado, el riesgo instantáneo de morir entre una persona blanca y una asiática o isleña es prácticamente igual.

En la variable **Grado**, se puede ver como a medida que se tiene un grado superior el *hazard ratio* aumenta. Aumenta de manera tan grande que se puede observar como una persona que ha padecido un cáncer de grado IV tiene un riesgo instantáneo 3,2 veces superior a una persona que lo ha padecido de grado I.

Por último, se puede observar como para la variable edad, una persona 20 años superior a otra tendrá un riesgo instantáneo de morir $\mathbf{\exp(0{,}009 * 20) = 1{,}2}$ veces superior.

Bondad de ajuste del modelo

A continuación en el Gráfico 4.10, se observarán los gráficos de residuos de Schoenfeld para cada covariable del modelo y así se observará si se cumple la premisa de riesgos proporcionales. Para que la premisa se cumpla, la curva del estimador en función del tiempo debe ser una línea recta con pendiente 0.

Posteriormente en la Tabla 4.6, se presentarán los valores de los coefiecientes de correlación entre el tiempo de supervivencia transformado y los residuos de Schoenfeld escalados (rho), una prueba chi-cuadrado y su p-valor. Para aceptar la premisa de riesgos proporcionales el p-valor de la prueba debería tener un valor superior a nuestro nivel de confianza, este caso no pasará en casi ninguna covariable debido a la gran cantidad de datos, por lo tanto, no se tendrá en cuenta este valor.

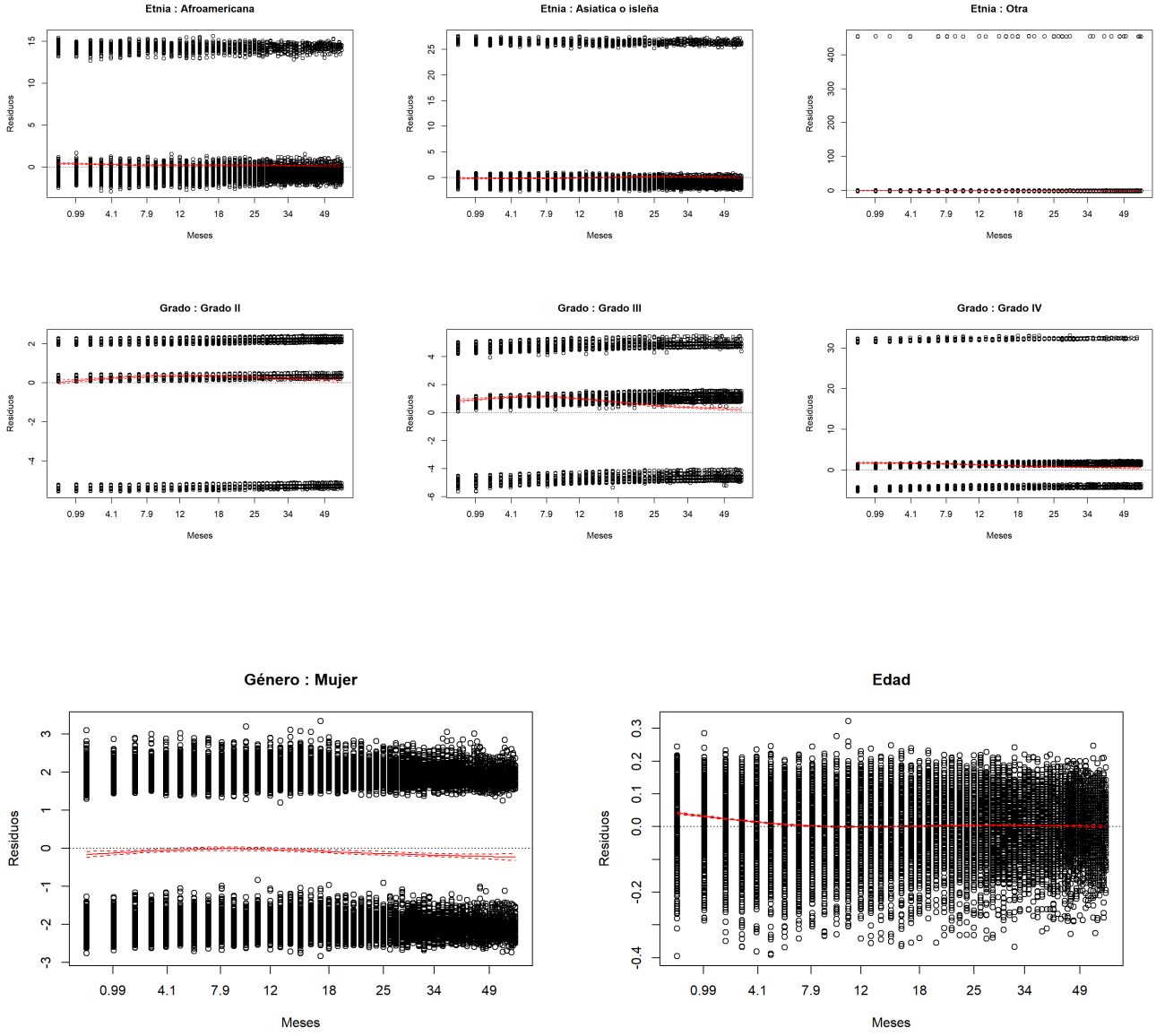


Figura 4.10: Residuos de Schoenfeld para la base de datos 1975-1984

El Gráfico 4.10 muestra como la premisa de riesgos proporcionales se cumple para todas las covariables excepto para *Grado III* y *Edad*, en estas covariables se puede observar como la línea roja no sigue la pendiente 0. Los resultados obtenidos para las covariables *Grado III* y *Edad* no se puede considerar que se cumplan a lo largo del tiempo.

Variable	**Categoría**	**rho**	$\boldsymbol{\chi^2}$	**P-valor**
Género	Mujer	-0.019	33.95	$\mathbf{5{,}67e^{-09}}$
Etnia	Afroamericana	-0.018	0.48	0.49
	Asiática o isleña	0.021	26.09	$\mathbf{3{,}25e^{-07}}$
	Otras	-0.003	0.171	0.68
Grado	Grado II	0.007	251.53	$\mathbf{1{,}207e^{-56}}$
	Grado III	-0.086	343.97	$\mathbf{8{,}72e^{-77}}$
	Grado IV	-0.073	110.54	$\mathbf{7{,}47e^{-26}}$
Edad		-0.097	291.93	$\mathbf{1{,}89e^{-65}}$
GLOBAL		NA	820.06	$\mathbf{9{,}78e^{-172}}$

Tabla 4.6: Valores de rho, chi cuadrado y p-valor, para validar el modelo de Cox de la base de datos 1975-1984.

Los valores de rho confirman lo que se ha visto gráficamente, se puede ver como los valores más lejanos del 0 son precisamente los de las covariables que se han comentado anteriormente que no siguen una línea horizontal sobre el eje 0.

4.3.2. Período 2004-2011

Igual que se ha hecho en el modelo de Cox para la base de datos de 1975 a 1984, se observarán los gráficos de los residuos de martingalas para comprobar si la forma continua es la más óptima para la variable **Edad**.

A causa de la gran cantidad de *missings* causados por las variables Seguro Médico y Tamaño del tumor, se han hecho dos gráficos distintos (Gráfico 4.11a y Gráfico 4.11b), ya que el primero, el cual contiene todos los datos, no nos da información. Se ha decidido mostrarlo para contrastarlo con el gráfico sin *missings*.

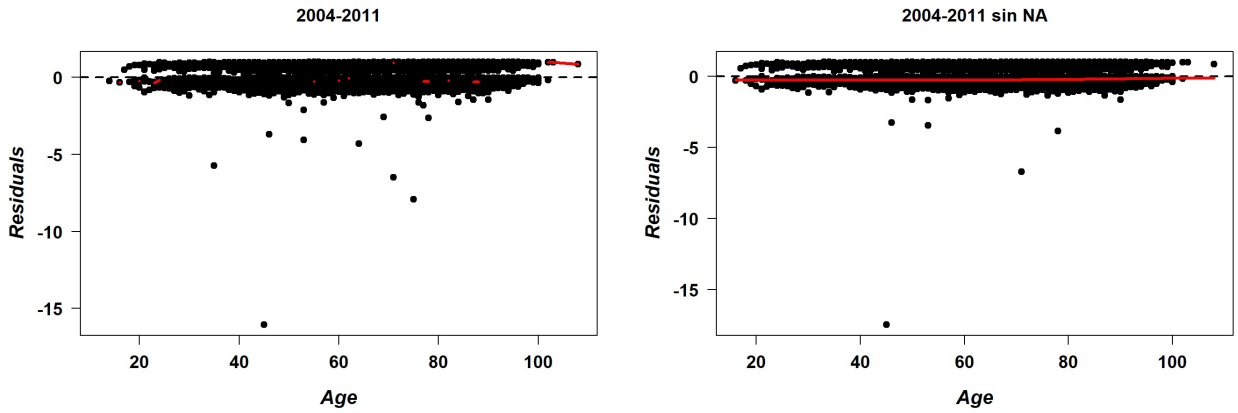


Figura 4.11: Gráficas de los residuos de Martingalas para la variable **Edad** con *missings* y sin *missings* en la base de datos 2004-2011

Observando el Gráfico 4.11b, no cabe ninguna duda que la forma de la variable **Edad** es la óptima. La línea roja sigue de forma horizontal el valor 0 de los residuos para todos los años de edades que hay en los datos.

Para la base de datos correspondiente al intervalo de años entre 2004 y 2011, se incluirán 12 covariables. Las categorías de referencia serán *Hombre*, *Blanca*, *Grado I* y *No asegurado*.

El modelo de riesgos proporcionales tendrá la siguiente expresión:

$$\boldsymbol{\lambda(t|\mathrm{X}) = \lambda_0(t)\exp\left\{\sum_{i}^{12}\beta_i \mathrm{X}_i\right\}}$$

Variable	Categoría	Coeficiente	HR	0.95 IC	P-valor
Género	Mujer	-0.078	0.925	**[0,888 , 0,965]**	**0,0003**
Etnia	Afroamericana	0.331	1.392	**[1,308 , 1,481]**	$\boldsymbol{< 2e^{-16}}$
	Asiática o isleña	-0.059	0.943	**[0,879 , 1,011]**	**0,098**
	Otras	-0.046	0.955	**[0,777 , 1,174]**	**0,660143**
Grado	Grado II	0.437	1.547	**[1,413 , 1,695]**	$\boldsymbol{< 2e^{-16}}$
	Grado III	1.119	3.062	**[2,783 , 3,37]**	$\boldsymbol{< 2e^{-16}}$
	Grado IV	1.261	3.528	**[3,091 , 4,027]**	$\boldsymbol{< 2e^{-16}}$
Seguro	Seguro público	-0.064	0.938	**[0,816 , 1,077]**	**0,363**
Médico	Asegurado	-0.458	0.633	**[0,557 , 0,719]**	$\boldsymbol{1{,}82e^{-12}}$
	Desconocido	-0.36	0.698	**[0,575 , 0,847]**	**0,0003**

Variable	Categoría	Coeficiente	HR	0.95 IC	P-valor
Tamaño del tumor		0.003	1.003	**[1,003 , 1,004]**	$< 2e^{-16}$
Edad		0.017	1.017	**[1,015 , 1,018]**	$< 2e^{-16}$

Tabla 4.7: Ajuste del modelo de Cox para el período 2004-2011.

En el modelo para los pacientes diagnosticados entre 2004 y 2011 ha habido un total de 9185 eventos y se han incluido 33160 pacientes.

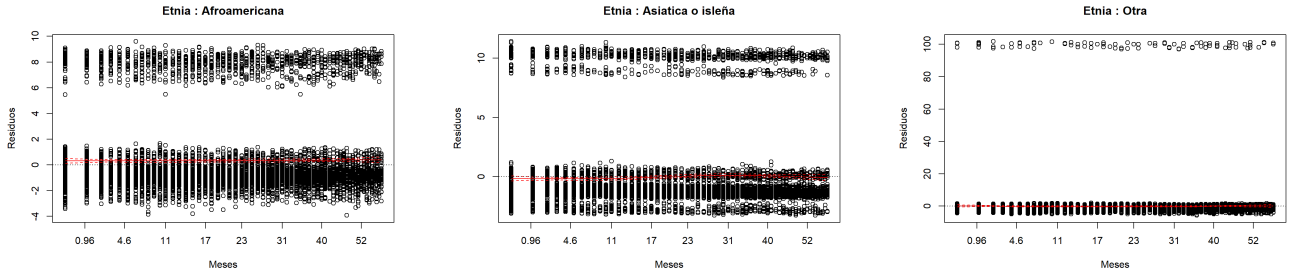
La Tabla 4.7 muestra como covariables son estadísticamente significativas excepto las categorías *Asiática o isleña* y *Otras* de la variable **Etnia** y la categoría *Seguro Público* de la variable **Seguro médico**. Coincide que de estas 3 categorías había menos datos.

Observando el riesgo instantáneo de las distintas categorías de la variable **Seguro Médico**, se ve claramente como el riesgo instantáneo es mayor para las personas no aseguradas, aunque las personas con un seguro público tienen un riesgo de fallecer por cáncer colorrectal prácticamente igual. Por otro lado, una persona no asegurada tiene un riesgo instantáneo de morir 1.58 veces superior a una persona asegurada.

Se puede ver como para la variable **Tamaño del tumor**, una persona con un tumor 30 mm superior a otra tendrá un riesgo instantáneo de morir 1.094 veces superior.

Validación

En el Gráfico 4.12, se van a presentar los gráficos de residuos de Schoenfeld para las covariables del modelo de Cox para la base de datos de los pacientes diagnosticados entre 2004 y 2011. Posteriormente, se presentarán los valores de rho, una prueba chi-cuadrado y su p-valor.



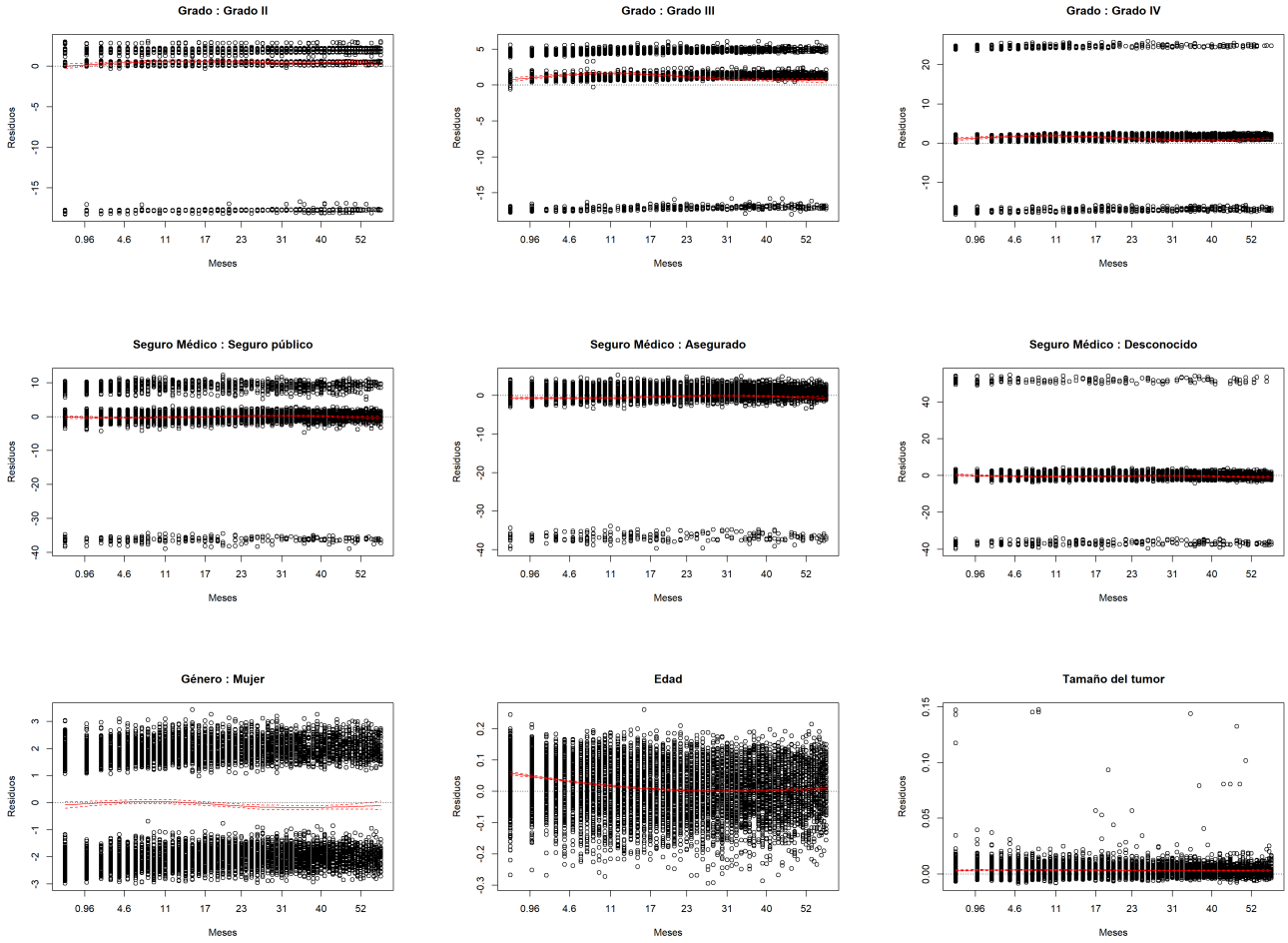


Figura 4.12: Residuos de Schoenfeld para la base de datos 2004-2011

Si se observan los gráficos de los residuos de schoenfeld (Gráfico 4.12), se puede ver como todas las covariables excepto **Edad** siguen una línea horizontal sobre el eje 0, es decir, cumplen la premisa de riesgos proporcionales y se verifica la hipótesis de que son constantes en el tiempo.

Variable	**Categoría**	**rho**	χ^2	**P-valor**
Género	Mujer	-0.028	42.38	$\mathbf{7{,}52e^{-11}}$
Etnia	Afroamericana	0.008	10.36	$\mathbf{1{,}29e^{-03}}$
	Asiática o isleña	0.025	13.42	$\mathbf{2{,}49e^{-04}}$
	Otras	0.0006	0.77	0.38

Variable	Categoría	rho	χ^2	P-valor
Grado	Grado II	0.017	164.52	$\mathbf{1{,}17e^{-37}}$
	Grado III	-0.041	130	$\mathbf{4{,}09e^{-30}}$
	Grado IV	-0.043	36.86	$\mathbf{1{,}27e^{-09}}$
Seguro Médico	Seguro público	0.015	0.61	0.435
	Asegurado	0.025	0.082	0.774
	Desconocido	-0.014	15.31	$\mathbf{9{,}15e^{-05}}$
Tamaño del tumor		-0.028	3.52	0.061
Edad		-0.184	419.61	$\mathbf{2{,}97e^{-93}}$
GLOBAL		NA	616.08	$\mathbf{3{,}9e^{-124}}$

Tabla 4.8: Valores de rho, chi cuadrado y p-valor, para validar el modelo de Cox de la base de datos 2004-2011.

En la Tabla 4.8, el valor de rho para la covariable **Edad** este es mucho más lejano al 0 que todos los demás. Es coherente con lo comentado anteriormente.

4.3.3. Comparación de ambos períodos

Observando los resultados de las distintas categorías de los modelos para cada base de datos. Se deberá prestar atención a ciertos cambios que se comentarán a continuación.

Para empezar, se puede observar como el riesgo instantáneo de muerte por cáncer colorrectal para una persona afroamericana en comparación a una persona blanca se ha visto incrementando en los pacientes que han sido diagnosticados entre 2004 y 2011. Este riesgo instantáneo ha pasado de 1.28 en la Tabla 4.5 (página 56) a 1.39 en la Tabla 4.7 (página 60).

Otro aspecto que vale la pena destacar, ha sido que los pacientes que han sido diagnosticados con un tumor de grado distinto a grado I, han visto un incremento en su riesgo instantáneo en comparación a dicho grado de tumor. Por ejemplo, las personas que han sido diagnosticadas con un tumor de grado II, en la base de datos más reciente (Tabla 4.7, página 60), tienen un riesgo instantáneo de morir 1.55 veces superior a los de grado I, en la base de datos del 1975-1984 (Tabla 4.5, página 56) era 1.29. Para los diagnosticados con un tumor de grado III ha sido aún mayor dicho incremento, han pasado de 2.18 a 3.06.

Capítulo V

Conclusiones

En este trabajo se han presentado tres métodos distintos para analizar la supervivencia de los pacientes de cáncer colorrectal en Estados Unidos. En primer lugar, se ha hecho un análisis no paramétrico que nos ha ayudado a ver cómo se comportan las curvas de supervivencia. Seguidamente se han ajustado el modelo de vida acelerada (paramétrico) y el modelo de Cox de riesgos proporcionales (semiparamétrico).

Al utilizar el modelo de Cox se deberá cumplir la premisa de riesgos proporcionales, lo cual suele ser inconveniente, sobretodo en los estudios con datos de enfermedades de larga duración. El modelo paramétrico de vida acelerada en muchos casos puede ser más eficiente y potente que el modelo de Cox, siempre y cuando los datos se ajusten bien a la distribución paramétrica elegida. Además, el modelo de vida acelerada, si es correcto, es más útil para estimar probabilidades de supervivencia y generar predicciones. Por otro lado, el modelo de Cox no necesita que los datos se ajusten a ninguna distribución paramétrica y permite el cálculo del *hazard ratio* (en el modelo paramétrico solo lo permite si se ajusta a una distribución Weibull).

El objetivo principal del estudio fue analizar los factores asociados a muerte por cáncer colorrectal en Estados Unidos a través del modelo de riesgos proporcionales y el modelo de vida acelerada. Durante el análisis de los datos se estableció un nuevo objetivo: la comparación entre los pacientes diagnosticados entre los años 1975 y 1984 y los diagnosticados entre 2004 y 2011. Las variables seleccionadas para ambos períodos han sido **Género**, **Etnia**, **Grado** y **Edad**, y para la segunda base de datos (2004-2011) también se han considerado **Tamaño del tumor** y **Seguro Médico**. Al ver los resultados del Gráfico 2.6 (página 31), se introdujo la interacción entre las variables **Seguro Médico** y **Etnia** en los dos modelos para la base de datos de 2004 a 2011, al ver que ni en el modelo de vida acelerada ni en el modelo de Cox las covariables de la interacción eran significativas, se decidió prescindir de esta.

Para comparar los riesgos de supervivencia entre las distintas covariables, se ha usado el factor de aceleración para los modelos de vida acelerada y el riesgo instantáneo para el modelo de riesgos proporcionales.

Observando las validaciones de los distintos modelos que se han hecho, se considera que el modelo de vida acelerada ajustado por una distribución log-normal es el que mejor se ajusta a los datos de los pacientes diagnosticados entre 1975 y 1984 (ver Gráfico 4.7b en página 51). La desventaja de un modelo ajustado por esta distribución es que la única medida del tamaño de efecto es el factor de aceleración. Para la base de

datos de años más recientes se considera mejor el modelo de Cox, ya que el modelo paramétrico que mejor se ajusta no tiene un buen un ajuste (ver Gráfico 4.8a en página 53) y en cambio, en el modelo de riesgos proporcionales, 11 de 12 covariables cumplen la premisa de riesgos proporcionales (ver Gráfico 4.12 en página 61 y Tabla 4.8 en página 62).

Tras observar los resultados obtenidos en el análisis no paramétrico, lo primero que vale la pena destacar es la diferencia de supervivencia para cada covariable en las dos bases de datos. En el período más reciente todas las covariables tienen una mayor supervivencia. Como se explica en el primer capítulo, en los últimos años se ha invertido en investigación y desarrollo para encontrar tratamientos más eficaces contra el cáncer, por lo tanto, el resultado obtenido parece ser razonable.

Los resultados obtenidos en los dos modelos mencionados anteriormente apuntan en la misma dirección para las dos bases de datos. Se ha podido observar como para las dos bases de datos las personas afroamericanas han sido las que han tenido un mayor riesgo de muerte por cáncer colorrectal a lo largo de los cinco primeros años después del diagnóstico. Además, este riesgo se ha visto aumentando en la segunda base de datos, es decir, entre 2004 y 2011 las personas afroamericanas tenían un mayor riesgo de muerte que las personas de otras etnias en comparación a las afroamericanas que fueron diagnosticadas entre 1975 y 1984. Por otro lado, como cabía esperar, las personas de mayor edad o las personas que fueron diagnosticadas con un cáncer de grado alto (III o IV) tienen un mayor riesgo de muerte por cáncer colorrectal.

Por lo que respecta a las variables **Seguro Médico** y **Tamaño del tumor** en la base de datos del período reciente, se han obtenido unos resultados esperados: las personas con un tumor de mayor tamaño tienen más riesgo de morir y las personas con seguro médico tienen mayor supervivencia. Lo que cabe destacar es que el riesgo de morir a lo largo de cinco años entre una persona sin seguro médico y una persona con seguro público es prácticamente igual, hecho que parece sugerir una baja calidad del seguro médico público estadounidense. Observando el Gráfico 2.6 (página 31) se puede comprobar como la etnia que más carece de seguro médico es la afroamericana, esto puede ayudar a entender los resultados comentados anteriormente. Por otro lado, es curioso que las personas asiáticas o isleñas no hayan tenido un mayor riesgo de muerte que las personas de etnia blanca (han tenido un riesgo casi igual e incluso un poco menor), ya que tienen muchas más personas sin seguro médico.

Pensando en futuros trabajos, sería interesante la recogida de nuevas variables socioeconómicas, por ejemplo, el tipo de seguro médico o la renta de cada paciente, ya que en Estados Unidos hay muchos tratamientos que no son cubiertos por ciertas aseguradoras. Debido a los resultados obtenidos para las personas de etnia asiática o isleña sería interesante la recogida de datos relacionados con el estilo de vida (alimentación, deporte, etc.), ya que como se comenta en el Capítulo I los factores de riesgo del cáncer colorrectal están fuertemente relacionados con el tipo de alimentación y el estilo de vida sedentario.

Bibliografia

[1] *¿Qué es el Cáncer colorrectal?* URL: https://www.cancer.org/es/cancer/cancer-de-colon-o-recto/acerca/que-es-cancer-de-colon-o-recto.html.

[2] *¿Qué es el Cáncer?* URL: https://www.cancer.gov/espanol/cancer/naturaleza/que-es.

[3] J Martin Bland y Douglas G Altman. "The logrank test". En: *BMJ (Clinical research ed.)* 328 (mayo de 2004). DOI: 10.1136/bmj.328.7447.1073.

[4] *Cáncer de colon y recto.* URL: https://seom.org/info-sobre-el-cancer/colon-recto?showall=1.

[5] J W Chapman. "Innovative estimation of survival using log-normal survival modelling on ACCENT database". En: *British journal of cancer* 108 (mar. de 2013), pág. 20. DOI: 10.1038/bjc.2013.34.

[6] D. R. Cox. "Regression Models and Life-Tables". En: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2 (1972), págs. 187-220. ISSN: 00359246. DOI: 10.2307/2985181. URL: http://www.jstor.org/stable/2985181.

[7] Paolo Di Lorenzo. *usmap: US Maps Including Alaska and Hawaii.* R package version 0.5.0. 2019. URL: https://CRAN.R-project.org/package=usmap.

[8] William Dudley, Rita Wickham y Nicholas Coombs. "An Introduction to Survival Statistics: Kaplan-Meier Analysis". En: *Journal of the Advanced Practitioner in Oncology* 7 (feb. de 2016). DOI: 10.6004/jadpro.2016.7.1.8.

[9] *Factores de riesgo del cáncer colorrectal.* URL: https://www.cancer.org/es/cancer/cancer-de-colon-o-recto/causas-riesgos-prevencion/factores-de-riesgo.html.

[10] Guadalupe Gómez, Olga Julià y Klaus Langohr. *Análisis de Supervivencia.*

[11] E. J GUMBEL. *Statistics of extremes.* 1958.

[12] E. L. Kaplan y Paul Meier. "Nonparametric Estimation from Incomplete Observations". En: *Journal of the American Statistical Association* 53.282 (1958), págs. 457-481. ISSN: 01621459. URL: http://www.jstor.org/stable/2281868.

[13] Elahe Khaksar y col. "Cox Regression and Parametric Models: Comparison of How They Determine Factors Influencing Survival of Patients with Non-Small Cell Lung Carcinoma". En: *Asian Pacific journal of cancer prevention : APJCP* 18 (dic. de 2017), págs. 3389-3393. DOI: 10.22034/APJCP.2017.18.12.3389.

[14] *Las Cifras del Cáncer en España 2020.* 2020. URL: https://seom.org/seomcms/images/stories/recursos/Cifras_del_cancer_2020.pdf.

[15] Isaac Subirana, Héctor Sanz y Joan Vila. "Building Bivariate Tables: The compareGroups Package for R". En: *Journal of Statistical Software* 57.12 (2014), págs. 1-16. URL: http://www.jstatsoft.org/v57/i12/.

[16] *Tasas de supervivencia por etapas para el cáncer colorrectal*. URL: https://www.cancer.org/es/cancer/cancer-de-colon-o-recto/deteccion-diagnostico-clasificacion-por-etapas/tasas-de-supervivencia.html.

[17] Terry M Therneau. *A Package for Survival Analysis in S*. version 2.38. 2015. URL: https://CRAN.R-project.org/package=survival.

[18] *Tratamiento del cáncer de colon según la etapa*. URL: https://www.cancer.org/es/cancer/cancer-de-colon-o-recto/tratamiento/por-etapas-colon.html.

[19] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

[20] Hadley Wickham, Jim Hester y Romain Francois. *readr: Read Rectangular Text Data*. R package version 1.3.1. 2018. URL: https://CRAN.R-project.org/package=readr.

Apéndice A

Código R

```
## Paquetes a usar

```{r warning=FALSE}

if (!(require(gridExtra))){
install.packages("gridExtra",repos = "http://cran.us.r-project.org")
}
if (!(require(readr))){
install.packages("readr",repos = "http://cran.us.r-project.org")
}
if (!(require(model4you))){
install.packages("model4you",repos = "http://cran.us.r-project.org")
}

if (!(require(SurvRegCensCov))){
install.packages("SurvRegCensCov",repos = "http://cran.us.r-project.org")
}
if (!(require(survival))){
install.packages("survival",repos = "http://cran.us.r-project.org")
}
if (!(require(survMisc))){
install.packages("survMisc",repos = "http://cran.us.r-project.org")
}
if (!(require(knitr))){
install.packages("knitr",repos = "http://cran.us.r-project.org")
}
if (!(require(ggplot2))){
install.packages("ggplot2",repos = "http://cran.us.r-project.org")
```
```

```
}
if (!(require(compareGroups))){
install.packages("compareGroups",repos = "http://cran.us.r-project.org")
}
if (!(require(KMsurv))){
install.packages("KMsurv",repos = "http://cran.us.r-project.org")
}
if (!(require(survminer))){
install.packages("survminer",repos = "http://cran.us.r-project.org")
}
if (!(require(flexsurv))){
install.packages("flexsurv",repos = "http://cran.us.r-project.org")
}
if (!(require(actuar))){
install.packages("actuar",repos = "http://cran.us.r-project.org")
}
if (!(require(dplyr))){
install.packages("dplyr",repos = "http://cran.us.r-project.org")
}
if (!(require(ggpubr))){
install.packages("ggpubr",repos = "http://cran.us.r-project.org")
}
if (!(require(magrittr))){
install.packages("magrittr",repos = "http://cran.us.r-project.org")
}
if (!(require(nph))){
install.packages("nph",repos = "http://cran.us.r-project.org")
}

if (!(require(sf))){
install.packages("sf",repos = "http://cran.us.r-project.org")
}
if (!(require(raster))){
install.packages("raster",repos = "http://cran.us.r-project.org")
}

if (!(require(spData))){
install.packages("spData",repos = "http://cran.us.r-project.org")
}

if (!(require(spDataLarge))){
install.packages("spDataLarge",repos = "http://cran.us.r-project.org")
```

```
}

if (!(require(tmap))){
install.packages("tmap",repos = "http://cran.us.r-project.org")
}# for static and interactive maps


if (!(require(leaflet))){
install.packages("leaflet",repos = "http://cran.us.r-project.org")
}# for interactive maps


if (!(require(mapview))){
install.packages("mapview",repos = "http://cran.us.r-project.org")
}# for interactive maps

if (!(require(shiny))){
install.packages("shiny",repos = "http://cran.us.r-project.org")
} # for web applications

if (!(require(usmap))){
install.packages("usmap",repos = "http://cran.us.r-project.org")
}
if (!(require(rms))){
install.packages("rms",repos = "http://cran.us.r-project.org")
}
```

```

## Lectura de datos

```{r warning=FALSE}
path_dades_1975_2016 <- 'C:/Users/Victor/Desktop/TFG/Dades_1975_2016/SEER_1975_2016_TEXTDATA/incidence/y
start_col <- c(1,9,19,20,24,25,28,35,37,39,43,47,48,52,53,57,58,59,60,61,64,66,68,69,71,73,86,88,92,93,9
end_col <-   c(8,18,19,21,24,27,31,36,38,42,46,47,51,52,56,57,58,59,60,63,65,67,68,70,72,85,87,91,92,93,
colname_dades_1975_2016 <- c("PATIENT_ID","REGISTRY_ID","MARITAL_STAT","ETHNICITY","SEX","AGE","BIRTH_YE
dades <- read_fwf(file=path_dades_1975_2016,col_positions=fwf_positions(start=start_col,end=end_col,col_
dades <- data.frame(dades)
variablesfinals <- c("PATIENT_ID","REGISTRY_ID","MARITAL_STAT","SEX","AGE","YEAR_DIAG","PRIMARY_SITE","C
dades <- dades[ 1:nrow(dades),variablesfinals]
head(dades)
```
```

Transformación de Variables e inclusión de Missings (NA).

```{r}
dades$AGE <- as.numeric(dades$AGE)

dades$MARITAL_STAT <- factor(dades$MARITAL_STAT)
levels(dades$MARITAL_STAT) <- c("Soltero","Casado","Separado","Divorciado","Viudo","No casado","Desconoc

dades$SEX <- factor(dades$SEX)
levels(dades$SEX) <- c("Hombre","Mujer")

dades$SURVIV_MONTHS <- as.numeric(dades$SURVIV_MONTHS)

dades$GRADE <- factor(dades$GRADE)
levels(dades$GRADE) <- c("Grado I","Grado II","Grado III","Grado IV","T-Cell","B-Cell","Null Cell","NK C

dades$DIAG_CONFIRM <- factor(dades$DIAG_CONFIRM)
levels(dades$DIAG_CONFIRM) <- c("Pos Histology","Pos Cytology","Pos Histo Plus","Pos Microscopic Confirm

dades$TYPE_REPORT_SOURCE <- factor(dades$TYPE_REPORT_SOURCE)
levels(dades$TYPE_REPORT_SOURCE) <- c("Hospital inpatient","Radiation Treatment","Lab Only ","Private Me



dades$REGIONAL_NODES_POS <- as.numeric(dades$REGIONAL_NODES_POS)
dades$CS_TUMOR_SIZE <- as.numeric(dades$CS_TUMOR_SIZE)

dades$SURGERY <- factor(dades$SURGERY)
levels(dades$SURGERY) <- c("Cirugía realizada","No recomendada","Contraindicada","Paciente murió antes",

dades$BEHAV_RECOD_ANALYS <- factor(dades$BEHAV_RECOD_ANALYS)
levels(dades$BEHAV_RECOD_ANALYS) <- c("Bening","Borderline Malignancy","In situ","Malignant","Only Malig

dades$RACE_RECODE_Y <- factor(dades$RACE_RECODE_Y)
levels(dades$RACE_RECODE_Y) <- c("Blanca","Afroamericana","American Indian/Alaska Native","Asiatica o is

dades$FIRST_MALIG_PRIM_IND <- factor(dades$FIRST_MALIG_PRIM_IND)
levels(dades$FIRST_MALIG_PRIM_IND) <- c("No","Yes")

dades$VITAL_STATUS_RECOD <- factor(dades$VITAL_STATUS_RECOD)
levels(dades$VITAL_STATUS_RECOD) <- c("Dead","Alive")

dades$INSURANCE_RECOD <- factor(dades$INSURANCE_RECOD)
```

```
levels(dades$INSURANCE_RECOD) <- c("No asegurado","Seguro público","Asegurado","Insured/No specific","De

dades$PRIMARY_SITE <- factor(dades$PRIMARY_SITE)

dades$SEER_OTH_DEATH_CAUSE <- factor(dades$SEER_OTH_DEATH_CAUSE)
dades$SURVIV_MONTHS_FLAG <- factor(dades$SURVIV_MONTHS_FLAG)

dades$AGE[dades$AGE==999] <- NA
dades$SURVIV_MONTHS <- as.numeric(dades$SURVIV_MONTHS)
dades$SURVIV_MONTHS[dades$SURVIV_MONTHS==9999] <-NA
dades$SURVIV_MONTHS[dades$SURVIV_MONTHS==0] <- 0.5
dades$SEER_DEATH_CAUSE[dades$SEER_DEATH_CAUSE ==9] <-NA
dades$REGIONAL_NODES_POS[dades$REGIONAL_NODES_POS >=90] <-NA # el 90, 95, 97, 98 y 99 tienen codificaci

table(dades$CS_TUMOR_SIZE)
dades$CS_TUMOR_SIZE[dades$CS_TUMOR_SIZE >989] <-NA
dades$CS_TUMOR_SIZE[dades$CS_TUMOR_SIZE ==888] <-NA# el 989,990,991,992,993,994,995,996,997,998,999,888

table(dades$SEER_DEATH_CAUSE)
dades$SEER_DEATH_CAUSE <- factor(dades$SEER_DEATH_CAUSE)

table(dades$SURVIV_MONTHS_FLAG)
dades$SURVIV_MONTHS_FLAG <- as.numeric(dades$SURVIV_MONTHS_FLAG)
dades$SURVIV_MONTHS_FLAG[dades$SURVIV_MONTHS_FLAG ==0] <- 0.5
dades$SURVIV_MONTHS_FLAG <- factor(dades$SURVIV_MONTHS_FLAG)

table(dades$SURGERY)
dades$SURGERY[dades$SURGERY== "Recommended, unknown if done"] <- "Desconocido"
dades$SURGERY[dades$SURGERY== "Unknown reason"] <- "Desconocido"
table(dades$SURGERY)
dades$SURGERY <- factor(dades$SURGERY)
table(dades$SURGERY)

table(dades$RACE_RECODE_Y)
dades$RACE_RECODE_Y[dades$RACE_RECODE_Y== "American Indian/Alaska Native"] <- "Otra"
dades$RACE_RECODE_Y <- factor(dades$RACE_RECODE_Y)
table(dades$RACE_RECODE_Y)

table(dades$INSURANCE_RECOD)
dades$INSURANCE_RECOD[dades$INSURANCE_RECOD == "Insured/No specific"] <- "Asegurado"
dades$INSURANCE_RECOD <- factor(dades$INSURANCE_RECOD)
table(dades$INSURANCE_RECOD)
```

```
table(dades$GRADE)
dades$GRADE <- as.character(dades$GRADE)
dades$GRADE[dades$GRADE=="T-Cell"]<- "Desconocido"

dades$GRADE <- factor(dades$GRADE)
table(dades$GRADE)

table(dades$BEHAV_RECOD_ANALYS )
dades$BEHAV_RECOD_ANALYS <- as.character(dades$BEHAV_RECOD_ANALYS)
dades$BEHAV_RECOD_ANALYS[dades$BEHAV_RECOD_ANALYS!= "Borderline Malignancy"] <- "No Malignant"
dades$BEHAV_RECOD_ANALYS <- factor(dades$BEHAV_RECOD_ANALYS)
table(dades$BEHAV_RECOD_ANALYS )

dades <- dades[!is.na(dades$SEER_DEATH_CAUSE),]
dades <- dades[!is.na(dades$SURVIV_MONTHS),]

dades$REGISTRY_ID <- as.numeric(dades$REGISTRY_ID)

dades$REGISTRY_ID[dades$REGISTRY_ID==0000001501] <- "San Francisco-Oakland SMSA (1973)"
dades$REGISTRY_ID[dades$REGISTRY_ID==0000001502] <- "Connecticut (1973)"

dades$REGISTRY_ID[dades$REGISTRY_ID==0000001520] <- "Metropolitan Detroit (1973)"

dades$REGISTRY_ID[dades$REGISTRY_ID==0000001521] <- "Hawaii (1973)"
dades$REGISTRY_ID[dades$REGISTRY_ID==0000001522] <- "Iowa (1973)"
dades$REGISTRY_ID[dades$REGISTRY_ID==0000001523] <- "New Mexico (1973)"
dades$REGISTRY_ID[dades$REGISTRY_ID==0000001525] <- "Seattle (Puget Sound) (1974)"
dades$REGISTRY_ID[dades$REGISTRY_ID==0000001526] <- "Utah (1973)"
dades$REGISTRY_ID[dades$REGISTRY_ID==0000001527] <- "Metropolitan Atlanta (1975)"
dades$REGISTRY_ID[dades$REGISTRY_ID==0000001529] <- "Alaska"
dades$REGISTRY_ID[dades$REGISTRY_ID==0000001531] <- "San Jose-Monterey"
dades$REGISTRY_ID[dades$REGISTRY_ID==0000001535] <- "Los Angeles"
dades$REGISTRY_ID[dades$REGISTRY_ID==0000001537] <- "Rural Georgia"
dades$REGISTRY_ID[dades$REGISTRY_ID==0000001541] <- "Greater California (excluding SF, Los Angeles & SJ)
dades$REGISTRY_ID[dades$REGISTRY_ID==0000001542] <- "Kentucky"
dades$REGISTRY_ID[dades$REGISTRY_ID==0000001543] <-"Louisiana"
dades$REGISTRY_ID[dades$REGISTRY_ID==0000001544] <- "New Jersey"
dades$REGISTRY_ID[dades$REGISTRY_ID==0000001547] <- "Greater Georgia (excluding AT and RG)"

table(dades$SEER_DEATH_CAUSE)
head(dades[,19:20])
```

```
```
Dades separades

```{r}
dades1 <- dades[dades$YEAR_DIAG < 1985,-c(11,12,22)]
dades2 <- dades[dades$YEAR_DIAG > 2003,]
dades2 <- dades2[dades2$YEAR_DIAG < 2012,]
```
Nuevas variables supervivencia max. 5 años (60 meses)

```{r}
dades$survival5ys <- pmin(dades$SURVIV_MONTHS, 5 * 12)
dades$censura5ys <- ifelse(dades$SURVIV_MONTHS > 5 * 12, 0, dades$SEER_DEATH_CAUSE == 1)

dades1$survival5ys <- pmin(dades1$SURVIV_MONTHS, 5 * 12)
dades1$censura5ys <- ifelse(dades1$SURVIV_MONTHS > 5 * 12, 0, dades1$SEER_DEATH_CAUSE == 1)

dades2$survival5ys <- pmin(dades2$SURVIV_MONTHS, 5 * 12)
dades2$censura5ys <- ifelse(dades2$SURVIV_MONTHS > 5 * 12, 0, dades2$SEER_DEATH_CAUSE == 1)
```


Descriptiva Univariante
Variables Numericas
Resumen Numérico
Base de datos 1975-1984

```{r}
sum <- vector()
var <- vector()
for (i in 1:ncol(dades1)){
if(is.numeric(dades1[,colnames(dades1)[i]])){
c <- summary(dades1[,colnames(dades1)[i]],digits=4)
sd <- round(sd(dades1[,colnames(dades1)[i]],na.rm=T),4)

if(length(c)==6){
c <- c(as.vector(c),0,sd)
var <- c(var,colnames(dades1)[i])
sum <- c(sum,c)

}else{
c <- c(as.vector(c),sd)
var <- c(var,colnames(dades1)[i])
```

```
sum <- c(sum,c)
}
}
}

sum <- matrix(sum,nrow=length(var),byrow = T)
sum <- data.frame(sum,row.names = var)
colnames(sum) <- c(names(summary(dades1$AGE)),"StDv")
kable(sum[1:2,])

```
Base de datos 2004-2011
```{r}
sum <- vector()
var <- vector()
for (i in 1:ncol(dades2)){
if(is.numeric(dades2[,colnames(dades2)[i]])){
c <- summary(dades2[,colnames(dades2)[i]],digits=4)
sd <- round(sd(dades2[,colnames(dades2)[i]],na.rm=T),4)

if(length(c)==6){
c <- c(as.vector(c),0,sd)
var <- c(var,colnames(dades2)[i])
sum <- c(sum,c)

}else{
c <- c(as.vector(c),sd)
var <- c(var,colnames(dades2)[i])
sum <- c(sum,c)
}
}
}

sum <- matrix(sum,nrow=length(var),byrow = T)
sum <- data.frame(sum,row.names = var)
colnames(sum) <- c(names(summary(dades2$AGE)),"StDv")
kable(sum[1:4,])

```

Resumen Gráfico
Base de datos 1975-1984

```
```{r}
dades1num <- dades1[,var[1:2]]
for(i in 1:round(length(dades1num)/2,0)){
par(mfrow=c(2,2))
j <- i*2
if(j-1 <= length(dades1num) ){
hist(dades1num[,j-1],main=paste("Histograma de",colnames(dades1num)[j-1],sep=" "),ylab = "Frecuencia",xl
abline(v=mean(dades1num[,j-1]),col="red",lty=1)
abline(v=median(dades1num[,j-1]),col="blue",lty=2)
boxplot(dades1num[,j-1],main=paste("Boxplot de",colnames(dades1num)[j-1],sep=" "),ylab=colnames(dades1nu
}

if(j<= length(dades1num)){
hist(dades1num[,j],main=paste("Histograma de",colnames(dades1num)[j],sep=" "),ylab = "Freqüència",xlab=c
abline(v=mean(dades1num[,j]),col="red",lty=1)
abline(v=median(dades1num[,j]),col="blue",lty=2)
boxplot(dades1num[,j],main=paste("Boxplot de",colnames(dades1num)[j],sep=" "),ylab=colnames(dades1num)[j
}
}

g1 <- ggplot(dades1[!is.na(dades1$AGE),], aes(x =AGE)) +
geom_histogram(color="black",fill="blue") +labs(title="Histograma ",x="Edad", y = "Frecuencia")
g2 <- ggplot(dades1[!is.na(dades1$AGE),], aes(y=AGE)) +
geom_boxplot(outlier.colour="red", outlier.shape=8,
outlier.size=4,fill="blue")+ labs(title="Diagrama de cajas",y="Edad")

require(gridExtra)

grid.arrange(g1, g2, ncol=2, widths = c(1.5, 1),top="Variable Edad 1975-1984")
```
Base de datos 2004-2011

```{r}
dades2num <- dades2[,var[1:4]]
for(i in 1:round(length(dades2num)/2,0)){
par(mfrow=c(2,2))
j <- i*2
if(j-1 <= length(dades2num) ){
hist(dades2num[,j-1],main=paste("Histograma de",colnames(dades2num)[j-1],sep=" "),ylab = "Frecuencia",xl
abline(v=mean(dades2num[,j-1]),col="red",lty=1)
abline(v=median(dades2num[,j-1]),col="blue",lty=2)
```
```

```
boxplot(dades2num[,j-1],main=paste("Boxplot de",colnames(dades2num)[j-1],sep=" "),ylab=colnames(dades2nu
}

if(j<= length(dades2num)){
hist(dades2num[,j],main=paste("Histograma de",colnames(dades2num)[j],sep=" "),ylab = "Freqüència",xlab=c
abline(v=mean(dades2num[,j]),col="red",lty=1)
abline(v=median(dades2num[,j]),col="blue",lty=2)
boxplot(dades2num[,j],main=paste("Boxplot de",colnames(dades2num)[j],sep=" "),ylab=colnames(dades2num)[j
}
}

g1 <- ggplot(dades2[!is.na(dades2$CS_TUMOR_SIZE),], aes(x =CS_TUMOR_SIZE)) +
geom_histogram(color="black",fill="blue") +labs(title="Histograma ",x="Edad", y = "Frecuencia")
g2 <- ggplot(dades2[!is.na(dades2$CS_TUMOR_SIZE),], aes(y=CS_TUMOR_SIZE)) +
geom_boxplot(outlier.colour="red", outlier.shape=8,
outlier.size=4,fill="blue")+ labs(title="Diagrama de cajas",y="Edad")

grid.arrange(g1, g2, ncol=2, widths = c(1.5, 1),top="Variable Edad 2004-2011")
```

```

### Variables Categóricas
#### Resumen Numérico
##### Base de datos 1975-1984

```{r}
varcat <- vector()
for(i in 3:ncol(dades1)){
if(!is.numeric(dades1[,colnames(dades1)[i]]) && colnames(dades1)[i] != "PRIMARY_SITE" && colnames(dades1
{
varcat <- c(varcat,colnames(dades1)[i])
}
}
dades1cat <- dades1[,varcat]
p <- vector()
v <- vector()
t <- vector()
for (i in 1:length(dades1cat)){
for(j in 1:length(levels(dades1cat[,i]))){
p <- c(p,paste(colnames(dades1cat)[i],":",levels(dades1cat[,i])[j],sep=" "))
v <- c(v,table(dades1cat[,i])[j],paste(round((((table(dades1cat[,i])[j])/length(dades1cat[,i]))*100),2),

}
```
```

```
p <- c(p,paste("Total",colnames(dades1cat)[i],sep=" "))
v <- c(v,sum(table(dades1cat[,i])),paste(round((((sum(table(dades1cat[,i])))/sum(!is.na(dades1cat[,i])))
}
names(v) <- NULL
v <- matrix(v,ncol=2,byrow=T)
v <- data.frame("Numero de Observaciones"=v[,1],"Porcentaje de Observaciones"=v[,2],row.names=p )
kable(v)

dades1$MARITAL_STAT <- factor(dades1$MARITAL_STAT)
levels(dades1$MARITAL_STAT)
```

```
##### Base de datos 2004-2011

```{r}
varcat <- vector()
for(i in 3:ncol(dades2)){
if(!is.numeric(dades2[,colnames(dades2)[i]]) && colnames(dades2)[i] != "PRIMARY_SITE" && colnames(dades2
{
varcat <- c(varcat,colnames(dades2)[i])
}
}
dades2cat <- dades2[,varcat]
p <- vector()
v <- vector()
t <- vector()
for (i in 1:length(dades2cat)){
for(j in 1:length(levels(dades2cat[,i]))){
p <- c(p,paste(colnames(dades2cat)[i],":",levels(dades2cat[,i])[j],sep=" "))
v <- c(v,table(dades2cat[,i])[j],paste(round((((table(dades2cat[,i])[j])/length(dades2cat[,i]))*100),2),

}
p <- c(p,paste("Total",colnames(dades2cat)[i],sep=" "))
v <- c(v,sum(table(dades2cat[,i])),paste(round((((sum(table(dades2cat[,i])))/sum(!is.na(dades2cat[,i])))
}
names(v) <- NULL
v <- matrix(v,ncol=2,byrow=T)
v <- data.frame("Numero de Observaciones"=v[,1],"Porcentaje de Observaciones"=v[,2],row.names=p )
kable(v)
```

Resumen Gráfico
Base de datos 1975-1984
```

```{r}
for (i in 1:round(length(dades1cat)/4,0)){
par(mfrow=c(2,2))
j <- i*4
if(j-1 <= length(dades1cat)){
plot(dades1cat[,j-1],main=paste("Gráfico de Barras de",colnames(dades1cat)[j-1],sep=" "),ylab="Numero de
}
if(j-2 <= length(dades1cat)){
plot(dades1cat[,j-2],main=paste("Gráfico de Barras de",colnames(dades1cat)[j-2],sep=" "),ylab="Numero de
}
if(j-3 <= length(dades1cat)){
plot(dades1cat[,j-3],main=paste("Gráfico de Barras de",colnames(dades1cat)[j-3],sep=" "),ylab="Numero de
}
if(j <= length(dades1cat)){

plot(dades1cat[,j],main=paste("Gráfico de Barras de",colnames(dades1cat)[j],sep=" "),ylab="Numero de   c
}
}

g11 <- ggplot(dades1,aes(x=GRADE)) +
geom_bar(fill = "blue") + labs(y = "Frecuencia",x="Grado",title= "Diagrama de barras Grado")

g21 <- ggplot(dades1,aes(x=MARITAL_STAT)) +
geom_bar(fill = "blue") + labs(y = "Frecuencia",x="Estado Civil",title= "Diagrama de barras Estado Civil
g31 <- ggplot(dades1,aes(x=RACE_RECODE_Y)) +
geom_bar(fill = "blue") + labs(y = "Frecuencia",x="Etnia",title= "Diagrama de barras Etnia")
grid.arrange(g11,g21,g31,nrow=3)
g21
```
Base de datos 2004-2011
```{r}
for (i in 1:round(length(dades2cat)/4,0)){
par(mfrow=c(2,2))
j <- i*4
if(j-1 <= length(dades2cat)){
plot(dades2cat[,j-1],main=paste("Gráfico de Barras de",colnames(dades2cat)[j-1],sep=" "),ylab="Numero de
}
if(j-2 <= length(dades2cat)){
plot(dades2cat[,j-2],main=paste("Gráfico de Barras de",colnames(dades2cat)[j-2],sep=" "),ylab="Numero de
}
if(j-3 <= length(dades2cat)){
plot(dades2cat[,j-3],main=paste("Gráfico de Barras de",colnames(dades2cat)[j-3],sep=" "),ylab="Numero de
```

```
}
if(j <= length(dades2cat)){

plot(dades2cat[,j],main=paste("Gráfico de Barras de",colnames(dades2cat)[j],sep=" "),ylab="Numero de   c
}
}

g12 <- ggplot(dades2[complete.cases(dades2),],aes(x=INSURANCE_RECOD)) +
geom_bar(fill = "blue") + labs(y = "Frecuencia",x="Seguro Médico",title= "Diagrama de barras Seguro Médi

g22 <- ggplot(dades2,aes(x=MARITAL_STAT)) +
geom_bar(fill = "blue") + labs(y = "Frecuencia",x="Estado Civil",title= "Diagrama de barras Estado Civil
g32 <- ggplot(dades2,aes(x=RACE_RECODE_Y)) +
geom_bar(fill = "blue") + labs(y = "Frecuencia",x="Etnia",title= "Diagrama de barras Etnia")
grid.arrange(g21,g22,nrow=2)
grid.arrange(g31,g32,nrow=2)
g12
```
```
### Análisis de Missings

#### Base de datos 1975-1984

```{r}
na <- vector()
for(i in 1:length(colnames(dades1))){
na <- c(na,sum(is.na(dades1[,i])),paste(round(sum(is.na(dades1[,i]))/length(dades1[,i]),6)*100,"%",sep="
}
na <- matrix(na, ncol=2,byrow=T)
na <- data.frame("Numero NA"=na[,1],"Porcentaje NA"=na[,2],row.names=colnames(dades1))
kable(na)
```
Base de datos 2004-2011

```{r}
na <- vector()
for(i in 1:length(colnames(dades2))){
na <- c(na,sum(is.na(dades2[,i])),paste(round(sum(is.na(dades2[,i]))/length(dades2[,i]),6)*100,"%",sep="
}
na <- matrix(na, ncol=2,byrow=T)
na <- data.frame("Numero NA"=na[,1],"Porcentaje NA"=na[,2],row.names=colnames(dades2))
kable(na)
```
```

```
## Mapa de distribución de Datos
```{r}
data("statepop")
statepop <- statepop$abbr[statepop$abbr!="CT"]
statepop <- statepop[statepop!="HI"]
statepop <- statepop[statepop!="IA"]
statepop <- statepop[statepop!="GA"]
statepop <- statepop[statepop!="MI"]
statepop <- statepop[statepop!="NM"]
statepop <- statepop[statepop!="CA"]
statepop <- statepop[statepop!="WA"]
statepop <- statepop[statepop!="UT"]
mapdata <- data.frame("state"=c("CT","HI","IA","GA","MI","NM","CA","WA","UT",statepop),"SEER_NAMES"=c(na

mapdata1 <- data.frame("state"=c("CT","HI","IA","GA","MI","NM","CA","WA","UT",statepop),"SEER_NAMES"=c(n

mapdata2 <- data.frame("state"=c("CT","HI","IA","GA","MI","NM","CA","WA","UT",statepop),"SEER_NAMES"=c(n

plot_usmap(data = mapdata, values = "count", color = "black",labels=T) +
scale_fill_continuous(low = "white", high = "blue",name = "Casos en base de datos Total", label = scales

plot_usmap(data = mapdata1, values = "count", color = "black",labels=T) +
scale_fill_continuous(low = "white", high = "blue",name = "Casos en base de datos 1975-1984", label = sc

plot_usmap(data = mapdata2, values = "count", color = "black",labels=T) +
scale_fill_continuous(low = "white", high = "blue",name = "Casos en base de datos 2004-2011", label = sc

```
Anàlisi Bivariant
Base de datos 1975-1984
VITAL_STATUS_RECOD VS. SEX

```{r}

ggplot(data = dades1[!is.na(dades1$VITAL_STATUS_RECOD),c("VITAL_STATUS_RECOD","SEX")],aes(VITAL_STATUS_R

createTable(compareGroups(VITAL_STATUS_RECOD~SEX,data=dades1[!is.na(dades1$VITAL_STATUS_RECOD),c("VITAL_
```
VITAL_STATUS_RECOD VS. AGE

```{r}
```
```

```
ggplot(dades1[!is.na(dades1$VITAL_STATUS_RECOD),c("VITAL_STATUS_RECOD","AGE")], aes(x=AGE, fill=VITAL_ST
geom_histogram(position="identity", alpha=0.5)+
geom_density(alpha=0.6)
createTable(compareGroups(VITAL_STATUS_RECOD~AGE,data=dades1[!is.na(dades1$VITAL_STATUS_RECOD),c("VITAL_

```
AGE VS. GRADE

```{r}
ggplot(dades1[!is.na(dades1$GRADE),c("GRADE","AGE")], aes(x=AGE, fill=GRADE)) +
geom_histogram(position="identity", alpha=0.5)+
geom_density(alpha=0.6)

createTable(compareGroups(GRADE~AGE,data=dades1[!is.na(dades1$GRADE),c("GRADE","AGE")]),show.ratio = TRU
```
SURGERY VS. GRADE

```{r}
table(dades1$SURGERY)
d1 <- dades1[dades1$SURGERY!= "Desconocido",]
ggplot(data = d1[!is.na(d1$SURGERY),c("SURGERY","GRADE")],aes(GRADE))+ geom_bar(aes(fill = SURGERY), pos
createTable(compareGroups(SURGERY~GRADE,data=d1[!is.na(d1$SURGERY),c("SURGERY","GRADE")]),show.ratio = T
```
VITAL_STATUS_RECOD VS. GRADE

```{r}
ggplot(data = dades1[!is.na(dades1$VITAL_STATUS_RECOD),c("VITAL_STATUS_RECOD","GRADE")],aes(GRADE))+ geo

createTable(compareGroups(VITAL_STATUS_RECOD~GRADE,data=dades1[!is.na(dades1$VITAL_STATUS_RECOD),c("VITA
```
Base de datos 2004-2011
RACE_RECODE_Y VS. INSURANCE_RECOD

```{r}
kable(table(dades2[!is.na(dades2$INSURANCE_RECOD),c("RACE_RECODE_Y","INSURANCE_RECOD")]$RACE_RECODE_Y,da

ggplot(data = dades2[!is.na(dades2$INSURANCE_RECOD),c("RACE_RECODE_Y","INSURANCE_RECOD")],aes(RACE_RECOD

createTable(compareGroups(RACE_RECODE_Y~INSURANCE_RECOD,data=dades2[!is.na(dades2$INSURANCE_RECOD),c("RA
```
VITAL_STATUS_RECOD VS. SEX

```{r}
ggplot(data = dades2[!is.na(dades2$VITAL_STATUS_RECOD),c("VITAL_STATUS_RECOD","SEX")],aes(VITAL_STATUS_R

createTable(compareGroups(VITAL_STATUS_RECOD~SEX,data=dades2[!is.na(dades2$VITAL_STATUS_RECOD),c("VITAL_
```

VITAL_STATUS_RECOD VS. AGE

```{r}
ggplot(dades2[!is.na(dades2$VITAL_STATUS_RECOD),c("VITAL_STATUS_RECOD","AGE")], aes(x=AGE, fill=VITAL_ST
geom_histogram(position="identity", alpha=0.5)+
geom_density(alpha=0.6)
createTable(compareGroups(VITAL_STATUS_RECOD~AGE,data=dades2[!is.na(dades2$VITAL_STATUS_RECOD),c("VITAL_

```

SURGERY VS. INSURANCE_RECOD

```{r}
ggplot(data = dades2[!is.na(dades2$INSURANCE_RECOD),c("SURGERY","INSURANCE_RECOD")],aes(INSURANCE_RECOD)

createTable(compareGroups(INSURANCE_RECOD~SURGERY,data=dades2[!is.na(dades2$INSURANCE_RECOD),c("SURGERY"
```

AGE VS. GRADE

```{r}
ggplot(dades2[!is.na(dades2$GRADE),c("GRADE","AGE")], aes(x=AGE, fill=GRADE)) +
geom_histogram(position="identity", alpha=0.5)+
geom_density(alpha=0.6)

createTable(compareGroups(GRADE~AGE,data=dades2[!is.na(dades2$GRADE),c("GRADE","AGE")]),show.ratio = TRU
```

SURGERY VS. GRADE

```{r}
table(dades2$SURGERY)
d1 <- dades2[dades2$SURGERY!= "Desconocido",]
ggplot(data = d1[!is.na(d1$SURGERY),c("SURGERY","GRADE")],aes(GRADE))+ geom_bar(aes(fill = SURGERY), pos

createTable(compareGroups(SURGERY~GRADE,data=d1[!is.na(d1$SURGERY),c("SURGERY","GRADE")]),show.ratio = T
```

VITAL_STATUS_RECOD VS. GRADE

```{r}
```

```
ggplot(data = dades2[!is.na(dades2$VITAL_STATUS_RECOD),c("VITAL_STATUS_RECOD","GRADE")],aes(GRADE))+ geo

createTable(compareGroups(VITAL_STATUS_RECOD~GRADE,data=dades2[!is.na(dades2$VITAL_STATUS_RECOD),c("VITA
```

```

# Datos Hasta 5 años (60 Meses)
### Base de datos 1975-1984
#### Modelo de Supervivencia Global
```{r}

y1=Surv(dades1$survival5ys, dades1$censura5ys)
Kaplan_meier1 = survfit(y1~1)
plot(Kaplan_meier1,lty=c(1,2),col=c(1,2),
xlab="Tiempo de supervivencia en meses",ylab="Probabilidades de supervivencia", main="Curva de Supervive

```
KM SEX

```{r}
table(dades1$SEX)
Kaplan_meier_SEX1 = survfit(y1~SEX,data=dades1)
plot(Kaplan_meier_SEX1, col = c(1,2),lty = c(1,1),ylim = c(0,1),main ="Curva de Supervivencia del Modelo
legend("bottomleft",c("Hombre","Mujer"),col = c(1,2),lty = c(1,1))
```
KM RACE

```{r}
table(dades1$RACE_RECODE_Y)
Kaplan_meier_RACE1 = survfit(y1~RACE_RECODE_Y,data=dades1)
plot(Kaplan_meier_RACE1, col = c(1,2,3,4),lty = c(1,2,3,1),ylim = c(0,1),main ="Curva de Supervivencia d
legend("bottomleft",c("Blanca","Afroamericana","Asiatica o isleña","Otra"),col = c(1,2,3,4),lty = c(1,2,
```
KM MARITAL_STATUS

```{r}
table(dades1$MARITAL_STAT)

Kaplan_meier_MARITAL_STAT1 = survfit(y1~MARITAL_STAT,data=dades1)

plot(Kaplan_meier_MARITAL_STAT1, col = c(1,2,3,4,5,6,8),lty = c(1,2,3,1,2,3,1),ylim = c(0,1),main ="Curv
legend("bottomleft",c("Soltero",   "Casado", "Separado" , "Divorciado"   ,"Viudo"),col = c(1,2,3,4,5,6,8
```
```

```
```
KM GRADE

```{r}

table(dades1$GRADE)
Kaplan_meier_GRADE1 = survfit(y1~GRADE,data=dades1)

plot(Kaplan_meier_GRADE1, col = c(1,2,3,4,5),lty = c(1,2,3,1,2),ylim = c(0,1),main ="Curva de Superviven
legend("bottomleft",c("Grado I"  ,     "Grado II"    ,  "Grado III"   ,   "Grado IV"  ),col = c(1,2,3,4,

```
Base de datos 2004-2011
Modelo de Supervivencia Global

```{r}
y2=Surv(dades2$survival5ys, dades2$censura5ys)
Kaplan_meier2 = survfit(y2~1)
plot(Kaplan_meier2,lty=c(1,2),col=c(1,2),
xlab="Tiempo de supervivencia en meses",ylab="Probabilidades de supervivencia", main="Curva de Supervive
```
KM SEX

```{r}
table(dades2$SEX)
Kaplan_meier_SEX2 = survfit(y2~SEX,data=dades2)
plot(Kaplan_meier_SEX2, col = c(1,2),lty = c(1,1),ylim = c(0,1),
xlab="Tiempo de supervivencia en meses",ylab="Probabilidades de supervivencia", main="Curva de Supervive
legend("bottomleft",c("Hombre","Mujer"),col = c(1,2),lty = c(1,1))
```
KM RACE

```{r}
table(dades2$RACE_RECODE_Y)
Kaplan_meier_RACE2 = survfit(y2~RACE_RECODE_Y,data=dades2)
plot(Kaplan_meier_RACE2, col = c(1,2,3,4),lty = c(1,2,3,1),ylim = c(0,1),
xlab="Tiempo de supervivencia en meses",ylab="Probabilidades de supervivencia", main="Curva de Supervive
legend("bottomleft",c("Blanca","Afroamericana","Asiatica o isleña","Otra"),col = c(1,2,3,4),lty = c(1,2,
```
KM MARITAL_STATUS

```{r}
```

```
table(dades2$MARITAL_STAT)
Kaplan_meier_MARITAL_STAT2 = survfit(y2~MARITAL_STAT,data=dades2)
plot(Kaplan_meier_MARITAL_STAT2, col = c(1,2,3,4,5,6,8),lty = c(1,2,3,1,2,3,1),ylim = c(0,1),
xlab="Tiempo de supervivencia en meses",ylab="Probabilidades de supervivencia", main="Curva de Supervive
legend("bottomleft",c("Soltero",   "Casado", "Separado" , "Divorciado"   ,"Viudo" ,"No casado"),col = c(
```

```
#### KM GRADE

```{r}
table (dades2$GRADE)
Kaplan_meier_GRADE2 = survfit(y2~GRADE,data=dades2)
plot(Kaplan_meier_GRADE2, col = c(1,2,3,4,5),lty = c(1,2,3,1,2),ylim = c(0,1),
xlab="Tiempo de supervivencia en meses",ylab="Probabilidades de supervivencia", main="Curva de Supervive
legend("bottomleft",c("Grado I"  ,     "Grado II"    ,  "Grado III"   ,   "Grado IV"),col = c(1,2,3,4,5)
```

```
#### KM INSURANCE_RECOD

```{r}
table (dades2$INSURANCE_RECOD)
Kaplan_meier_INSURANCE_RECOD2 = survfit(y2~INSURANCE_RECOD,data=dades2)
plot(Kaplan_meier_INSURANCE_RECOD2, col = c(1,2,3,4),lty = c(1,2,3,1),ylim = c(0,1),
xlab="Tiempo de supervivencia en meses",ylab="Probabilidades de supervivencia", main="Curva de Supervive
legend("bottomleft",c("No asegurado","Seguro público","Asegurado", "Desconocido"),col = c(1,2,3,4),lty =
```

```
## Descriptiva Kaplan Meier Supervivencia (solo Funcion de Superv+Censura)
### Base de datos 1975-1984
```{r}
Kaplan_meier1
Kaplan_meier_SEX1
survdiff(y1~SEX,data=dades1)
Kaplan_meier_RACE1
survdiff(y1~RACE_RECODE_Y,data=dades1)
Kaplan_meier_MARITAL_STAT1
survdiff(y1~MARITAL_STAT,data=dades1)
Kaplan_meier_GRADE1
survdiff(y1~GRADE,data=dades1)
```

Base de datos 2004-2011
```{r}
Kaplan_meier2
Kaplan_meier_SEX2
```
```

```
survdiff(y2~SEX,data=dades2)
Kaplan_meier_RACE2
survdiff(y2~RACE_RECODE_Y,data=dades2)
Kaplan_meier_MARITAL_STAT2
survdiff(y2~MARITAL_STAT,data=dades2)
Kaplan_meier_GRADE2
survdiff(y2~GRADE,data=dades2)
Kaplan_meier_INSURANCE_RECOD2
survdiff(y2~INSURANCE_RECOD,data=dades2)
```

```
# Comparacion KM Entre Bases de datos 1975-1984 y 2004-2011
## Modelo de Supervivencia Global
```{r}
par(mfrow=c(1,2))
plot(Kaplan_meier1,lty=c(1,2),col=c(1,2),
xlab="Temps de supervivència en mesos",ylab="Probabilitats de supervivència", main="Corbes KM Superv Dad
plot(Kaplan_meier2,lty=c(1,2),col=c(1,2),
xlab="Temps de supervivència en mesos",ylab="Probabilitats de supervivència", main="Corbes KM Superv Dad
```
KM SEX
```{r}
par(mfrow=c(1,2))
plot(Kaplan_meier_SEX1, col = c(1,2),lty = c(1,1),ylim = c(0,1), main = "Corbes KM Superv 1975-1984")
legend("bottomleft",c("Hombre","Mujer"),col = c(1,2),lty = c(1,1))
plot(Kaplan_meier_SEX2, col = c(1,2),lty = c(1,1),ylim = c(0.3,1), main = "Corbes KM Superv 2004-2011")
legend("bottomleft",c("Hombre","Mujer"),col = c(1,2),lty = c(1,1))
```
KM RACE

```{r}
par(mfrow=c(1,2))
plot(Kaplan_meier_RACE1, col = c(1,2,3,4),lty = c(1,2,3,1),ylim = c(0,1),main = "Corbes KM Superv 1975-1
legend("bottomleft",c("Blanca","Afroamericana","Asiatica o isleña","Otra"),col = c(1,2,3,4),lty = c(1,2,
plot(Kaplan_meier_RACE2, col = c(1,2,3,4),lty = c(1,2,3,1),ylim = c(0,1),main = "Corbes KM Superv 1975-1
legend("bottomleft",c("Blanca","Afroamericana","Asiatica o isleña","Otra"),col = c(1,2,3,4),lty = c(1,2,
```

KM MARITAL_STATUS

```{r}
par(mfrow=c(1,2))
plot(Kaplan_meier_MARITAL_STAT1, col = c(1,2,3,4,5,6,8),lty = c(1,2,3,1,2,3,1),ylim = c(0,1),main = "Cor
```
```

```
legend("bottomleft",c("Soltero",   "Casado", "Separado" , "Divorciado"   ,"Viudo" ,"No casado",   "Desco
plot(Kaplan_meier_MARITAL_STAT2, col = c(1,2,3,4,5,6,8),lty = c(1,2,3,1,2,3,1),ylim = c(0,1),main = "Cor
legend("bottomleft",c("Soltero",   "Casado", "Separado" , "Divorciado"   ,"Viudo" ,"No casado",   "Desco
```

```
## KM GRADE
```{r}
par(mfrow=c(1,2))
plot(Kaplan_meier_GRADE1, col = c(1,2,3,4,5),lty = c(1,2,3,1,2),ylim = c(0,1),main = "Corbes KM Superv 1
legend("bottomleft",c("Grado I"  ,     "Grado II"    ,  "Grado III"   ,   "Grado IV"   ,      "Desconoci
plot(Kaplan_meier_GRADE2, col = c(1,2,3,4,5),lty = c(1,2,3,1,2),ylim = c(0,1),main = "Corbes KM Superv 2
legend("bottomleft",c("Grado I"  ,     "Grado II"    ,  "Grado III"   ,   "Grado IV"   ,      "Desconoci
```
Nuevos datos para modelos

```{r}
dades <- dades[dades$GRADE!="Desconocido",]
dades$GRADE <- factor(dades$GRADE)
dades <- dades[dades$MARITAL_STAT!="Desconocido",]
dades$MARITAL_STAT_n <- as.character(dades$MARITAL_STAT)
dades$MARITAL_STAT_n[dades$MARITAL_STAT=="Soltero"] <- "Single & Unmarried"
dades$MARITAL_STAT_n[dades$MARITAL_STAT=="No casado"] <- "Single & Unmarried"
dades$MARITAL_STAT_n[dades$MARITAL_STAT=="Separado"] <- "Separated & Divorced"
dades$MARITAL_STAT_n[dades$MARITAL_STAT=="Divorciado"] <- "Separated & Divorced"
dades$MARITAL_STAT_n <- factor(dades$MARITAL_STAT_n)
dades1 <- dades1[dades1$GRADE!="Desconocido",]
dades1$GRADE <- factor(dades1$GRADE)
dades1 <- dades1[dades1$MARITAL_STAT!="Desconocido",]
dades1$MARITAL_STAT_n <- as.character(dades1$MARITAL_STAT)
dades1$MARITAL_STAT_n[dades1$MARITAL_STAT=="Soltero"] <- "Single & Unmarried"
dades1$MARITAL_STAT_n[dades1$MARITAL_STAT=="No casado"] <- "Single & Unmarried"
dades1$MARITAL_STAT_n[dades1$MARITAL_STAT=="Separado"] <- "Separated & Divorced"
dades1$MARITAL_STAT_n[dades1$MARITAL_STAT=="Divorciado"] <- "Separated & Divorced"
dades1$MARITAL_STAT_n <- factor(dades1$MARITAL_STAT_n)
dades2 <- dades2[dades2$GRADE!="Desconocido",]
dades2$GRADE <- factor(dades2$GRADE)
dades2 <- dades2[dades2$MARITAL_STAT!="Desconocido",]
dades2$MARITAL_STAT_n <- as.character(dades2$MARITAL_STAT)
dades2$MARITAL_STAT_n[dades2$MARITAL_STAT=="Soltero"] <- "Single & Unmarried"
dades2$MARITAL_STAT_n[dades2$MARITAL_STAT=="No casado"] <- "Single & Unmarried"
dades2$MARITAL_STAT_n[dades2$MARITAL_STAT=="Separado"] <- "Separated & Divorced"
dades2$MARITAL_STAT_n[dades2$MARITAL_STAT=="Divorciado"] <- "Separated & Divorced"
```
```

```
dades2$MARITAL_STAT_n <- factor(dades2$MARITAL_STAT_n)
```
```
## Vida Acelerada
```{r}
mwei1 <- survreg(Surv(dades1$survival5ys, dades1$censura5ys) ~ SEX + RACE_RECODE_Y + GRADE  + AGE, data
mwei2 <- survreg(Surv(dades2$survival5ys, dades2$censura5ys) ~ SEX + RACE_RECODE_Y + GRADE + INSURANCE_
```
Base de datos Período 1975-1984
```{r}
summary(mwei1)
extractAIC(mwei1)
ConvertWeibull(mwei1,conf.level=0.95) #ETR
plot(residuals(mwei1))
abline(h=0,col="red")

# The residuals
# -------------
weipred <- predict(mwei1, type = "linear")
#weipred
resids <- (log(dades1$survival5ys) - weipred) / mwei1$scale
#resids <- resids[!is.na(resids)]

# The KM curve of the residuals
# -----------------------------
#windows(width = 10)
par(font = 2, font.lab = 4, font.axis = 2, las = 1, oma = c(0, 0, 1, 0),
mar = c(5, 5, 4, 2))
plot(survfit(Surv(resids,  dades1$censura5ys) ~ 1), xlab = "Residuos", lwd = 3,
ylab = expression(bold(hat(S)(t))), yaxs = "i")
title("Residuos del modelo weibull 1975 - 1984")

# Survival function of standard Gumbel distribution
survgumb <- function(x) {
exp(-exp(x))
}

# Adding the theoretical curve
curve(survgumb(x), from = min(resids), to = max(resids), col = 2, lwd = 3,
add = TRUE)
legend("bottomleft", c("Estimación KM", "95% - IC", "Distribución Gumbel estándar"),
col = c(1, 1, 2), lty = c(1, 2, 1), lwd = 3, bty = "n")
```
```
```
```

```
### Base de datos Período 2004-2011

'''{r}
summary(mwei2)
extractAIC(mwei2)
ConvertWeibull(mwei2,conf.level=0.95) #ETR
plot(residuals(mwei2))
abline(h=0,col="red")

# The residuals
# -------------
weipred <- predict(mwei2, type = "linear")
#weipred
resids <- (log(dades2$survival5ys) - weipred) / mwei2$scale
#resids <- resids[!is.na(resids)]

# The KM curve of the residuals
# -----------------------------
#windows(width = 10)
par(font = 2, font.lab = 4, font.axis = 2, las = 1, oma = c(0, 0, 1, 0),
mar = c(5, 5, 4, 2))
plot(survfit(Surv(resids, dades2$censura5ys) ~ 1), xlab = "Residuos", lwd = 3,
ylab = expression(bold(hat(S)(t))), yaxs = "i")
title("Residuos del modelo Weibull 2004-2011")

# Survival function of standard Gumbel distribution
survgumb <- function(x) {
exp(-exp(x))
}

# Adding the theoretical curve
curve(survgumb(x), from = min(resids[!is.na(resids)]), to = max(resids[!is.na(resids)]), col = 2, lwd =
add = TRUE)
legend("bottomleft", c("Estimación KM", "95% - IC", "Distribución Gumbel estándar"),
col = c(1, 1, 2), lty = c(1, 2, 1), lwd = 3, bty = "n")
'''
## Lognormal

'''{r}
mlogn1 <- survreg(Surv(dades1$survival5ys,dades1$censura5ys) ~ SEX + RACE_RECODE_Y + GRADE +  AGE, data

mlogn2 <- survreg(Surv(dades2$survival5ys,dades2$censura5ys) ~ SEX + RACE_RECODE_Y + GRADE + INSURANCE_R
```

```
```
Base de datos Período 1975-1984
```{r}
summary(mlogn1)
extractAIC(mlogn1)
ConvertWeibull(mlogn1,conf.level=0.95) #ETR
plot(residuals(mlogn1))
abline(h=0,col="red")

## Lognormal distribution
## ----------------------

# The residuals
# -------------
lnopred <- predict(mlogn1, type = "linear")
residsLN <- (log(dades1$survival5ys) - lnopred) / mlogn1$scale

# The KM curve of the residuals
# -----------------------------
#windows(width = 10)
par(font = 2, font.lab = 4, font.axis = 2, las = 1, oma = c(0, 0, 1, 0),
mar = c(5, 5, 4, 2))
plot(survfit(Surv(residsLN, dades1$censura5ys) ~ 1), xlab = "Residuos", lwd = 3,
ylab = expression(bold(hat(S)(t))), yaxs = "i")
title("Residuos del modelo lognormal 1975-1984")

# Adding the theoretical survival function
curve(pnorm(x, lower.tail = FALSE), from = min(residsLN[!is.na(residsLN)]), to = max(residsLN[!is.na(res
col = 2, lwd = 3, add = TRUE)
legend("bottomleft", c("Estimación KM", "95% - IC", "Distribución Normal estándar"),
col = c(1, 1, 2), lty = c(1, 2, 1), lwd = 3, bty = "n")
```
Base de datos Período 2004-2011

```{r}
summary(mlogn2)
extractAIC(mlogn2)
ConvertWeibull(mlogn2,conf.level=0.95) #ETR
plot(residuals(mlogn2))
abline(h=0,col="red")
## Lognormal distribution
## ----------------------
```

```
# The residuals
# -------------
lnopred <- predict(mlogn2, type = "linear")
residsLN <- (log(dades2$survival5ys) - lnopred) / mlogn2$scale

# The KM curve of the residuals
# -----------------------------
#windows(width = 10)
par(font = 2, font.lab = 4, font.axis = 2, las = 1, oma = c(0, 0, 1, 0),
mar = c(5, 5, 4, 2))
plot(survfit(Surv(residsLN, dades2$censura5ys) ~ 1), xlab = "Months", lwd = 3,
ylab = expression(bold(hat(S)(t))), yaxs = "i")
title("Residuos del modelo lognormal 2004-2011")

# Adding the theoretical survival function
curve(pnorm(x, lower.tail = FALSE), from = min(residsLN[!is.na(residsLN)]), to = max(residsLN[!is.na(res
col = 2, lwd = 3, add = TRUE)
legend("bottomleft", c("Estimación KM", "95% - IC", "Distribución Normal estándar"),
col = c(1, 1, 2), lty = c(1, 2, 1), lwd = 3, bty = "n")
```

```
## COX
```{r}
mcox1 <- coxph(Surv(dades1$survival5ys, dades1$censura5ys) ~ SEX + RACE_RECODE_Y + GRADE + AGE, data = c

mcox2 <- coxph(Surv(dades2$survival5ys, dades2$censura5ys) ~ SEX + RACE_RECODE_Y + GRADE + INSURANCE_REC
```

Función para calcular rho

```{r}
your_func <- function(fit, transform = "km", new_cox.zph = NULL) {
sresid <- resid(fit, "schoenfeld")
varnames <- names(fit$coefficients)
nvar <- length(varnames)
ndead <- length(sresid)/nvar
if (nvar == 1) {
times <- as.numeric(names(sresid))
} else {
times <- as.numeric(dimnames(sresid)[[1]])
}
```
```

```
if (is.character(transform)) {
tname <- transform
ttimes <- switch(transform, identity = times, rank = rank(times),
log = log(times), km = {
temp <- survfitKM(factor(rep(1, nrow(fit$y))),
fit$y, se.fit = FALSE)
t1 <- temp$surv[temp$n.event > 0]
t2 <- temp$n.event[temp$n.event > 0]
km <- rep(c(1, t1), c(t2, 0))
if (is.null(attr(sresid, "strata"))) 1 - km else (1 -
km[sort.list(sort.list(times))])
}, stop("Unrecognized transform"))
}
else {
tname <- deparse(substitute(transform))
if (length(tname) > 1)
tname <- "user"
ttimes <- transform(times)
}
xx <- ttimes - mean(ttimes)
r2 <- sresid %*% fit$var * ndead
test <- xx %*% r2
corel <- c(cor(xx, r2))
cbind(rho = c(corel,NA), new_cox.zph$table)
}
```

```
### Base de datos Período 1975-1984
```{r}
summary(mcox1)
#ggcoxdiagnostics(mcox1)

res.mcox1 <- cox.zph(mcox1,terms=F)
your_func(fit = mcox1, new_cox.zph = res.mcox1)
res.mcox1
summary(res.mcox1)

plot(res.mcox1[1], main= "Género : Mujer", xlab = "Meses",ylab="Residuos",col="red" )
abline(h=0,col="black",lty=3)

plot(res.mcox1[2], main= "Etnia : Afroamericana", xlab = "Meses",ylab="Residuos",col="red" )
abline(h=0,col="black",lty=3)
```
```

```
plot(res.mcox1[3], main= "Etnia : Asiático o isleño", xlab = "Meses",ylab="Residuos",col="red" )
abline(h=0,col="black",lty=3)

plot(res.mcox1[4], main= "Etnia : Otra", xlab = "Meses",ylab="Residuos",col="red" )
abline(h=0,col="black",lty=3)

plot(res.mcox1[5], main= "Grado : Grado II", xlab = "Meses",ylab="Residuos",col="red" )
abline(h=0,col="black",lty=3)

plot(res.mcox1[6], main= "Grado : Grado III", xlab = "Meses",ylab="Residuos",col="red" )
abline(h=0,col="black",lty=3)

plot(res.mcox1[7], main= "Grado : Grado IV", xlab = "Meses",ylab="Residuos",col="red" )
abline(h=0,col="black",lty=3)

plot(res.mcox1[8], main= "Edad", xlab = "Meses",ylab="Residuos",col="red" )
abline(h=0,col="black",lty=3)

‘‘‘
#### Gráfico de residuos de martingala con un modelo de Cox

‘‘‘{r}
#residuals(mcox1)## Martingale residuals
resids1 <- residuals(update(mcox1,  ~. - AGE))

## Checking for the linear assumption of the continuos covariates "age".
## ------------------------------------------------------------------
## Martingale residuals of the model WITHOUT variable Age

par(font = 2, font.lab = 4, font.axis = 2, las = 1, cex.lab = 1.3, cex.axis = 1.2)
plot(resids1 ~ dades1$AGE, xlab = "Age", ylab = "Residuals", pch = 19, main = "1975-1984")
abline(h = 0, lwd = 2, lty = 2)
lines(lowess(dades1[, "AGE"], resids1), lwd = 3, col="red")
‘‘‘
### Base de datos Período 2004-2011

‘‘‘{r}
summary(mcox2)
#ggcoxdiagnostics(mcox2)

res.mcox2 <- cox.zph(mcox2, terms=F)
#ggcoxzph(res.mcox2) # residuos de schoenfeld
```

```
your_func(fit = mcox2, new_cox.zph = res.mcox2)
res.mcox2
summary(res.mcox2)

plot(res.mcox2[1], main= "Género : Mujer", xlab = "Meses",ylab="Residuos",col="red" )
abline(h=0,col="black",lty=3)

plot(res.mcox2[2], main= "Etnia : Afroamericana", xlab = "Meses",ylab="Residuos",col="red" )
abline(h=0,col="black",lty=3)

plot(res.mcox2[3], main= "Etnia : Asiático o isleño", xlab = "Meses",ylab="Residuos",col="red" )
abline(h=0,col="black",lty=3)

plot(res.mcox2[4], main= "Etnia : Otra", xlab = "Meses",ylab="Residuos",col="red" )
abline(h=0,col="black",lty=3)

plot(res.mcox2[5], main= "Grado : Grado II", xlab = "Meses",ylab="Residuos",col="red" )
abline(h=0,col="black",lty=3)

plot(res.mcox2[6], main= "Grado : Grado III", xlab = "Meses",ylab="Residuos",col="red" )
abline(h=0,col="black",lty=3)

plot(res.mcox2[7], main= "Grado : Grado IV", xlab = "Meses",ylab="Residuos",col="red" )
abline(h=0,col="black",lty=3)

plot(res.mcox2[8], main= "Seguro Médico : Seguro público", xlab = "Meses",ylab="Residuos",col="red" )
abline(h=0,col="black",lty=3)

plot(res.mcox2[9], main= "Seguro Médico : Asegurado", xlab = "Meses",ylab="Residuos",col="red" )
abline(h=0,col="black",lty=3)

plot(res.mcox2[10], main= "Seguro Médico : Desconocido", xlab = "Meses",ylab="Residuos",col="red" )
abline(h=0,col="black",lty=3)

plot(res.mcox2[11], main= "Tamaño del tumor", xlab = "Meses",ylab="Residuos",col="red" )
abline(h=0,col="black",lty=3)
plot(res.mcox2[12], main= "Edad", xlab = "Meses",ylab="Residuos",col="red" )
abline(h=0,col="black",lty=3)

‘‘‘
#### Gráfico de residuos de martingala con un modelo de Cox
```

```r
#residuals(mcox2)## Martingale residuals
resids2 <- residuals(update(mcox2,  ~. - AGE)) #muchos NA??
#NA INFO
sum(is.na(dades2$AGE))
sum(is.na(resids2))
nrow(dades2)

## Checking for the linear assumption of the continuos covariates "age".
## ------------------------------------------------------------------
## Martingale residuals of the model WITHOUT variable Age
#windows(width=8)
par(font = 2, font.lab = 4, font.axis = 2, las = 1, cex.lab = 1.3, cex.axis = 1.2)
plot(resids2 ~ dades2$AGE, xlab = "Age", ylab = "Residuals", pch = 19, main= "2004-2011")
abline(h = 0, lwd = 2, lty = 2)
lw <- lowess(dades2[, "AGE"], resids2)
lines(lw[!is.na(lw)], lwd = 3, col="red")
```

```r
dades2sinna <- dades2[complete.cases(dades2),]
mcox2 <- coxph(Surv(dades2sinna$survival5ys,dades2sinna$censura5ys) ~ SEX + RACE_RECODE_Y + GRADE + INSU

resids2 <- residuals(update(mcox2,  ~. - AGE)) #muchos NA??
#NA INFO
sum(is.na(dades2$AGE))
sum(is.na(resids2))
nrow(dades2)
length(is.na(dades2$AGE))

## Checking for the linear assumption of the continuos covariates "age".
## ------------------------------------------------------------------
## Martingale residuals of the model WITHOUT variable Age
#windows(width=8)
par(font = 2, font.lab = 4, font.axis = 2, las = 1, cex.lab = 1.3, cex.axis = 1.2)
plot(resids2 ~ dades2sinna$AGE, xlab = "Age", ylab = "Residuals", pch = 19, main= "2004-2011 sin NA")
abline(h = 0, lwd = 2, lty = 2)
lw <- lowess(dades2sinna[, "AGE"], resids2)
lines(lw[!is.na(lw)], lwd = 3, col="red")
```