Universidad Nacional Autónoma de México

Facultad de Ciencias

Introducción al análisis de Riesgos Competitivos bajo el enfoque de la Función de Incidencia Acumulada (FIA) y su aplicación con R

T E S I S

QUE PARA OBTENER EL TÍTULO DE:
ACTUARIA

PRESENTA:
ANA PAMELA GUTIÉRREZ MARTÍNEZ

DIRECTOR DE TESIS:
M EN C. JOSÉ SALVADOR ZAMORA MUÑOZ

2016

Cd. Mx., 2016

Índice

Introducción

El análisis de supervivencia es una técnica que permite modelar el tiempo de falla. Por ejemplo, en el estudio de enfermedades crónicas o tratamientos muy agresivos, el tiempo hasta que ocurre la muerte del enfermo (tiempo de supervivencia) y su dependencia de la aplicación de distintos tratamientos, pero en otras enfermedades, el tiempo hasta la curación, o el tiempo hasta la aparición de la enfermedad. En procesos de control de calidad se estudia el lapso hasta que un cierto producto falla (tiempo de fallo), o el tiempo de espera hasta recibir un servicio (tiempo de espera), etc.

La necesidad de desarrollar tratamientos o programas para una enfermedad requiere un análisis de los resultados que sea específico para dicha enfermedad. Criterios de valoración como la insuficiencia cardiaca, la muerte debido a una enfermedad específica o el control de la enfermedad local en el cáncer pueden ser imposibles de observar debido a la aparición previa de un tipo de evento diferente (como la muerte por otra causa). El evento que dificulta o modifica la posibilidad de observar el de interés se denomina riesgo competitivo.

El análisis de riesgos competitivos tiene su origen en el año 1760, cuando David Bernoulli introdujo dicho concepto al realizar estudios con el fin de investigar el riesgo de morir a causa de la viruela u otra enfermedad. Este es un ejemplo clásico de riesgos competitivos, donde los individuos están expuestos a más de un riesgo de muerte y surge el interés por determinar cuál tiene mayor probabilidad de suceder. En los últimos años se han publicado trabajos relacionados con el análisis de riesgos competitivos en el campo de la medicina y la bioestadística.

La construcción de un modelo bajo el análisis de riesgos competitivos no es exclusiva de las áreas médicas y biológicas, se ha utilizado en otros campos, como: en el área de confiabilidad donde el interés se centra en determinar los posibles factores que causan la ruptura de algún componente, en las ciencias actuariales se ha utilizado bajo el nombre de decrementos múltiples, en demografía también se ha empleado para determinar la incidencia de diversos factores que definen comportamientos poblacionales, etc.

La característica distintiva de un modelo de riesgos competitivos es que el objeto de estudio se conforma por un par (T, C) donde $T > 0$, representa el tiempo transcurrido hasta el evento de interés, por otro lado, $C \in \{1, 2, \ldots, p\}$ representa el tipo de falla. Por tales características se requiere de un modelo conjunto para T y C, ya que se tienen p diferentes causas de falla y cada tipo se puede clasificar como un elemento perteneciente al conjunto $\{1, 2, \ldots, p\}$.

La distribución conjunta de (T, C) puede establecerse a través del riesgo de causa-específica. Es decir, el riesgo instantáneo de falla en un momento dado por una causa específica, entre todos los sujetos que se encuentran en riesgo en ese momento. La distribución conjunta puede determinarse a través de la función de incidencia acumulada, ésta representa la probabilidad de presentar la falla por una causa dada antes de un tiempo específico.

El objetivo de este trabajo es mostrar los conceptos principales del análisis de riesgos competitivos y su aplicación a un conjunto de datos mediante el programa `r.project`, utilizando la librería `cmprsk`. En el primer capítulo se describen las principales funciones que conforman al análisis de supervivencia: función de supervivencia, función de riesgo, función de riesgo acumulado, función de vida media residual y censura.

En el segundo capítulo se describen las funciones que se requieren para abordar un análisis de riesgos competitivos bajo el enfoque de la función de incidencia acumulada, como: función de riesgo de causa-específica, función de incidencia acumulada, función de incidencia acumulada de causa-específica. También se muestra la teoría que sustenta la construcción del paquete `cmprsk`, el cual se construyó a partir del Método de Gray, el cual introduce la teoría de r-muestras a través del análisis de covariables.

En el tercer capítulo se muestran dos de los métodos que a lo largo del tiempo se han utilizado para estimar riesgos competitivos. El complemento del estimador de Kaplan-Meier fue el primero que obtuvo resultados sobre datos analizados bajo el enfoque de riesgos competitivos. El segundo modelo que aborda este trabajo es, prueba de covariables, construído por el Dr. Gray en el año de 1999. Dicho método es la base del paquete `cmprsk` del programa estadístico `r.project`.

En el cuarto capítulo se describen las funciones que contiene el paquete `cmprsk` y las variables requeridas. En la segunda parte se analiza una base de datos (`ccup.csv`) con las funciones del paquete `cmprsk`: `crr`, `print.crr`, `predict.crr`, `plot.predict.crr`, `cuminc` y `plot.cuminc`. Se hace una explicación de la construcción de dichas funciones y lo que representan los resultados obtenidos. En la última sección se construye una simulación, la cual muestra la construcción de un análisis de riesgos competitivos a partir de una semilla y las funciones que se requieren para elaborar un análisis de riesgos competitivos.

Finalmente se presentan las conclusiones obtenidas sobre el análisis de riesgos competitivos a la base de datos `ccup.csv` y la simulación, así como la importancia y aportaciones que dicho enfoque tiene en diversos sectores.

Capítulo 1

Análisis de supervivencia

Introducción

El análisis de supervivencia tiene como objetivo modelar el tiempo transcurrido entre un punto inicial y un evento de interés, denominado tiempo de falla.

El evento de interés o falla se caracteriza por sólo ocurrir una única vez a cualquier individuo. Se le llamó análisis de supervivencia porque las primeras aplicaciones se realizaron en el sector salud, utilizando como evento de interés la muerte de los pacientes.

El tiempo de supervivencia se define como el tiempo transcurrido desde la entrada al estudio o estado inicial hasta el estado final o la ocurrencia del evento de interés.

Las funciones comúnmente utilizadas en el análisis de supervivencia, son: la de supervivencia, la de densidad de probabilidad, la de riesgo, la de riesgo acumulado y la de vida media residual.

1.1. Tiempo de falla

En muchos análisis y tratamientos médicos el tiempo juega un papel muy importante porque se convierte en el objeto de estudio, su significado depende de lo que el investigador esté analizando, por ejemplo, sería de interés: conocer el tiempo que un paciente permanece en el hospital después de recibir algún tratamiento, el tiempo que transcurre antes de presentar complicaciones posteriores a la aplicación de un tratamiento, tiempo en que tarda un medicamento en surtir efecto; tiempo que tarda en presentarse la muerte, etc. A estos diversos escenarios se les conoce como eventos de interés. El tiempo de falla puede manejarse en diferentes unidades: horas, días, semanas, meses, años, etc.

En el análisis de supervivencia se trabaja con el tiempo y éste tiene un papel fundamental, por ello, es muy importante la fecha de inicio del seguimiento de los pacientes. En muchos casos los pacientes entran al estudio a lo largo de un periodo, lo que implica que no todos los individuos tienen la misma fecha de inicio, esto quiere decir que hay entradas escalonadas. Por lo tanto, el tiempo de falla para cada individuo usualmente se mide a partir de su fecha de entrada.

El tiempo en el que sucede el evento de interés puede ser continuo o discreto (más adelante se describen las diferencias). Algo que caracteriza a la variable tiempo utilizada en el análisis de supervivencia es que es una variable aleatoria no negativa.

Para estudiar el tiempo de falla bajo el análisis de supervivencia, existen diversas funciones:

- Función de supervivencia
- Función de densidad de probabilidad
- Función de riesgo
- Función de riesgo acumulado
- Función de vida media residual

1.2. Censura

La censura es el fenómeno que ocurre cuando el valor de una observación sólo se conoce parcialmente. Los datos censurados aparecen con frecuencia cuando la variable de interés es el tiempo de supervivencia[1].

Las observaciones dentro del análisis de supervivencia se estudian a través del tiempo. Cuando el seguimiento término antes de que suceda el evento de interés o de completar el periodo de observación, se habla de censura. Cuando los tiempos de supervivencia no se conocen con exactitud, los datos también se consideran censuras.

El análisis de supervivencia, en general, busca obtener resultados de tratamientos clínicos aplicados a grupos de pacientes, con el fin de conocer el tiempo que tarda en reaccionar al tratamiento o lo que tardará en desarrollarse una enfermedad, entre otros intereses médicos. Por ello, el seguimiento viene definido por una fecha de inicio y de fin, que específica el tiempo de observación. Las fechas de inicio y fin pueden ser distintas para cada individuo, porque los pacientes incluídos en el estudio se incorporan en momentos diferentes. También existe el caso en que todos los sujetos comienzan y terminan al mismo tiempo el estudio.

En las observaciones censuradas (incompletas) el evento de interés no se ha producido, ya sea, porque el estudio finalizó antes de la aparición del evento de interés, el paciente decide abandonar y no participar más en el estudio, se pierde al paciente por cambio en el lugar de residencia, muerte no relacionada con la investigación, etc.

Existen diversos tipos de censura, éstas dependen de qué mecanismo las produzca. Las que se consideran son las siguientes:

- **Tipo I.** Ocurre cuando los individuos entran al estudio en diferente tiempo y el fin del estudio está predeterminado por el investigador.

[1]Muñoz Rodríguez José Elías, *"Modelación de Datos Espaciales Censurados"*. Facultad de Matemáticas Universidad de Guanajuato, México, 2005.

- **Tipo II.** Hay dependencia en el tamaño de muestra (n) y las fallas que se observan. El tiempo de observación término cuando r de los n individuos presentaron el evento de interés. Donde r es un número entero positivo determinado previamente por el investigador, tal que, $r < n$.
- **Aleatoria.** Surge cuando los sujetos salen del estudio por razones no controladas por el investigador y no presentaron la falla.
- **Por la izquierda.** Sucede cuando el evento de interés ocurre antes de que el investigador inicie la observación de los sujetos.
- **Por intervalo.** Este tipo de censura se presenta cuando se tiene un estudio de supervivencia donde el seguimiento del estado de los sujetos se realiza periódicamente, ocasionando que la falla sólo pueda conocerse entre dos periodos de revisión.

1.3. Función de supervivencia

La función de supervivencia, que se denota como $S(t)$, describe los fenómenos de tiempo-evento. Con ella se obtiene la probabilidad de que un individuo bajo estudio no experimente el evento de interés antes de un momento dado. Es decir, para una variable aleatoria T no negativa con función de distribución $F(t)$ y función de densidad de probabilidad $f(t)$:

$$\begin{aligned} S(t) &= Pr(T > t) \\ &= Pr(\text{probabilidad de que un sujeto sobreviva más allá del tiempo } t) \end{aligned}$$

También se puede expresar de la siguiente manera.

$$\begin{aligned} S(t) &= 1 - F(t) \\ &= 1 - Pr(T \leq t) \\ &= 1 - Pr(\text{probabilidad de presentar la falla antes del tiempo, } t) \end{aligned}$$

Por lo tanto, $S(t)$ es una función no creciente:

$$S(0) = 1 \quad \text{y} \quad S(t) = 0 \quad \text{cuando} \quad t \to \infty$$

Esto quiere decir, que la probabilidad de sobrevivir al menos al tiempo cero es igual a 1 y de sobrevivir un tiempo infinito es igual a 0.

Cuando T, es una variable aleatoria continua, la función de supervivencia es la integral de la función de densidad de probabilidad.

$$S(t) = P(T > t) = \int_t^{\infty} f(u)du$$

En el caso en que T, es una variable aleatoria discreta, se toman valores t_j, con $j = 1, 2, \ldots, k$ y función de masa de probabilidad:

$$f(t_j) = Pr(T = t_j)$$

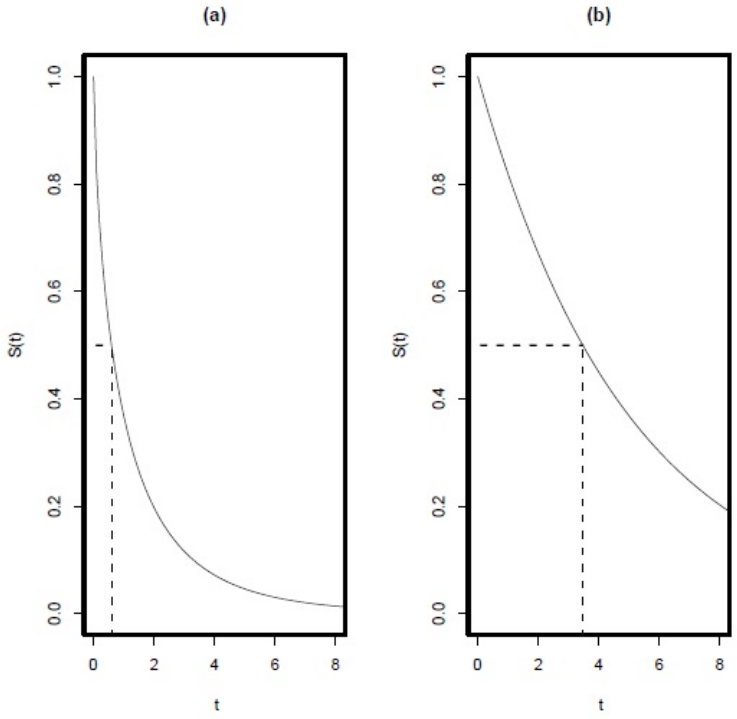
Con $(t_1 < \ldots < t_k)$. En este caso la función de supervivencia está dada por:

$$S(t) = P(T > t) = \sum_{t_j > t} f(t_j).$$

La representación gráfica de $S(t)$ muestra el comportamiento de la función y se le llama curva de supervivencia. Cumple las siguientes propiedades:

1. Monótona no creciente.
2. Igual a uno al inicio del seguimiento, o al tiempo cero.
3. Igual a cero cuando el tiempo tiende a infinito.

Al gráficar esta curva se obtiene un análisis de gran utilidad e importancia en la práctica. Es normal comparar dos o más curvas de supervivencia con el fin de estudiar y entender el comportamiento que se tiene entre ellas a lo largo del tiempo.



Gráfica 3.1: Curvas de supervivencia.

Una curva de supervivencia como la 3.1 (a) representa una tasa baja de supervivencia. Las curvas de supervivencia planas o graduales representan tasas altas de supervivencia.

1.4. Función densidad de probabilidad

El tiempo de supervivencia T, tiene una función de probabilidad como cualquier variable aleatoria. La función densidad de probabilidad de una variable aleatoria continua, describe la probabilidad relativa que tomará.

En el caso continuo se expresa como:

$$\begin{aligned} S(t) &= P(T > t) \\ &= \int_t^{\infty} f(u)du \\ f(t) &= -\frac{dS(t)}{dt} \end{aligned}$$

- Donde $f(u)du$, puede ser (de manera aproximada) la probabilidad de que un evento pueda ocurrir al tiempo, u.
- $f(u)$, una función no negativa con área bajo ella igual a uno.

A partir de la función de densidad se puede encontrar la proporción de individuos que cae en cualquier intervalo de tiempo y el pico de frecuencia más alto de fallas.

La función de densidad discreta toma valores positivos únicamente en los puntos del recorrido y se interpreta como la probabilidad de que la variable tome ese valor.

$$f(t_j) = Pr(T = t_j) \qquad \text{con} \qquad j = 1, 2, \ldots, k$$

1.5. Función de riesgo

Representa la tasa instantánea del evento para un individuo bajo estudio que ha llegado al tiempo t, sin experimentar el evento de interés. Cuando el evento de interés es la muerte, ésta es la tasa instantánea de muerte para un individuo que ha sobrevivido al tiempo t. Esta función es muy importante dentro del análisis de supervivencia, porque muestra cómo la tasa instantánea del evento varía con respecto al tiempo.

En términos matemáticos, la función de riesgo es la probabilidad condicional de que un evento ocurra dentro de un estrecho intervalo de tiempo, es decir, entre $(t, t+\delta t)$, dado que no hubo ningún evento hasta el tiempo t. Donde δt debe ser muy pequeño. Se expresa como:

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{Pr(t < T \leq t + \delta t | T > t)}{\delta t} \right\}$$

Esta expresión se puede modificar algebraicamente para obtener una forma más sencilla:

$$\begin{aligned} h(t) &= \lim_{\delta t \to 0} \left\{ \frac{Pr(t < T \leq t + \delta t)}{\delta t Pr(T > t)} \right\} \\ &= \frac{1}{Pr(T > t)} \lim_{\delta t \to 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\} \\ &= -\frac{1}{Pr(T > t)} * \frac{dF(t)}{\delta t} \\ &= \frac{f(t)}{S(t)} \end{aligned}$$

La función de supervivencia se puede expresar en términos de la de riesgo, en el caso continuo es igual a:

$$h(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)} = -\frac{d}{dt}\{\log S(t)\}$$

Existe una función relacionada con la de riesgo, conocida como riesgo acumulado y se define como:

$$H(t) = \int_0^t h(x)dx$$

La función de supervivencia, función de riesgo y función de riesgo acumulado, pueden relacionarse entre si:

$$S(t) = \exp\left[-H(t)\right] = \exp\left[-\int_0^t h(x)dx\right]$$

En el caso discreto, la función de riesgo se define para valores t_j, y proporciona la probabilidad condicional de falla al tiempo $t = t_j$, dado que el individuo estaba vivo antes de t_j, por lo tanto se tiene que:

$$\begin{aligned} h(t_j) &= Pr(T = t_j | T \geq t_j) \\ &= \frac{Pr(T = t_j)}{Pr(T \geq t_j)} \\ &= \frac{f(t_j)}{S(t_{j^{-1}})} \end{aligned}$$

Donde $t_{j^{-1}}$ corresponde a un instante antes de t_j.

$$\begin{aligned} Pr(T \geq t_j) &= 1 - Pr(T < t_j) \\ &= S(t_{j^{-1}}) \\ &\neq S(t_j) \end{aligned}$$

La función de riesgo describe la forma en que cambia la tasa instantánea de la ocurrencia de un evento de interés al paso del tiempo, la única restricción que tiene es que debe ser no negativa, $h(t) \geq 0$. Puede crecer, decrecer, permanecer constante, etc.

1.6. Función de riesgo acumulado

Esta función se define como $H(t)$, y se encarga de acumular el riesgo hasta el tiempo t. En el caso continuo y discreto significa lo mismo. La función es no decreciente y conforme se incrementa se puede obtener información del comportamiento del riesgo mientras transcurre el tiempo, t.

En el caso continuo se expresa como:

$$H(t) = \int_0^t h(u)du$$

Para el caso discreto se pueden utilizar dos opciones. En el primer caso:

$$H(t) = \sum_{t_j \leq t} h(t_j)$$

Y en el segundo,

$$H(t) = -\sum_{t_j < t} \ln\left[1 - h(t_j)\right]$$

1.7. Función de vida media residual

Esta función mide la esperanza del tiempo de vida restante o el tiempo esperado antes de la ocurrencia del evento de interés.

Se define como,

$$mrl(t) = E(T - t | T > t)$$

Para el caso continuo la función de vida media residual al tiempo t, se define como:

$$\begin{aligned} mrl &= E(T-t|T>t) \\ &= \int_0^\infty (u-t) f(u|u>t)du \\ &= \int_0^\infty (u-t)\frac{f(u)}{S(t)} I_{(t,\infty)}(u)du \\ &= \int_t^\infty \frac{(u-t)f(u)}{S(t)}du \\ &= \frac{\int_t^\infty (u-t)f(u)du}{S(t)} \\ &= \frac{\int_t^\infty S(u)du}{S(t)} \end{aligned}$$

Se puede observar que la función de vida media residual es el área bajo la curva de supervivencia a la derecha de t, dividida entre $S(t)$.

En el caso discreto la función de vida media residual está dada por:

$$\begin{aligned} mrl &= E(T-t|T>t) \\ &= \sum_{r=0}^{\infty} (t_r - t)\, P\left[T = t_r | T > t\right] \\ &= \sum_{r=0}^{\infty} \frac{(t_r - t)\, P\left[T = t_r, T > t\right]}{S(t)} \end{aligned}$$

Sea $t_i \leq t < t_{i+1}$ para alguna $i = 0, 1, 2, ...$

$$mrl = \frac{(t_{i+1} - t)S(t_i) + \sum_{j>i+1} (t_{j+1} - t_j)S(t_j)}{S(t)} \tag{1.1}$$

Para $t_i \leq t < t_{i+1}$

Capítulo 2

Riesgos competitivos

Introducción

La teoría que se ha utilizado en el análisis de supervivencia se desarrolló con el fin de estudiar la falla como único evento de interés. Dentro de este análisis sólo se permite observar esta variable (falla). La realidad ha mostrado que en la mayoría de los estudios, los pacientes durante el periodo de observación, antes de llegar al evento de interés, pueden presentar diversos eventos que provocan la muerte, la aceleran o la evitan. En la práctica habitual, la aparición de otro evento se toma como censura. Los investigadores observaron que tratar como censura otro tipo de evento, en algunos casos, influye en la probabilidad de supervivencia del análisis. Es por ello que la falla diferente a la de interés se nombró como riesgo competitivo.

Un escenario de riesgos competitivos se presenta cuando el paciente se encuentra expuesto a p tipos de fallas incluyendo la de interés y la ocurrencia de una impide inmediatamente que suceda el evento de interés u otra.

Por ejemplo, un grupo de pacientes diagnósticado con una enfermedad cardíaca, es observado para determinar el tiempo que transcurre a partir de la detección de la enfermedad hasta la muerte por infarto al miocardio. Si al final del estudio hubo registro de muertes causadas por infarto al miocardio, implíca que el evento de interés se observó y todos aquellos que sobrevivieron más allá del periodo de observación del análisis son censuras.

Por las características del grupo para analizar los resultados se utilizaría el estimador Kaplan-Meier[1], porque las observaciones por infarto al miocardio son el evento de interés. Sin embargo, en la vida real algunos pacientes pudieron haber muerto debido a otra falla, antes de experimentar un infarto al miocardio. Esto es una situación de riesgos competitivos, porque la muerte se presenta por diversas causas incluyendo la de interés. Bajo el análisis de riesgos competitivos otras causas de muerte no se consideran censuras, su observación no es incompleta. A partir de esto, surge la pregunta ¿cómo analizar los riesgos competitivos?

El modelo más utilizado ha sido el complemento del estimador de Kaplan-Meier $(1 - KM)$, por su uso recurrente en el análisis de una sola falla y su fácil construcción, pero al emplear esta técnica se obtienen resultados sesgados y la interpretación puede ser no clara. Es por ello que la estimación de la probabilidad del evento de interés, debe calcularse con herramientas específicas para el análisis de riegos competitivos. En los últimos años se han utilizado: la función de incidencia acumulada (introducida por Kalbfleisch y Prentice), de riesgo-específico y de incidencia acumulada de causa-específica.

Los resultados obtenidos en el análisis de riesgos competitivos se componen de dos términos (T, C). Donde T representa el tiempo de falla, y C, es la causa que la originó. El tiempo de falla T, es una variable aleatoria y C, puede tomar un número pequeño fijo de valores entre $\{1, \ldots, p\}$. Dentro del marco básico de probabilidad, es una distribución bivariada en donde uno de los componentes es discreto, en este caso es C, y el otro continuo, T.

En este trabajo se presenta cómo analizar riesgos competitivos utilizando la función de incidencia acumulada (FIA) en un modelo, identificando los factores asociados a una causa-específica de falla. Los individuos que presentan cualquier falla distinta a la de interés, no se consideran datos censurados.

[1] Apéndice A.

2.1. Función de riesgo de causa-específica

Para el modelo de riesgos competitivos se define una tasa instantánea de falla de causa-específica. También se determina la función de riesgo al tiempo t, debido a la causa j, en presencia de todas las otras fallas:

La función de riesgo de causa-específica representa la tasa instantánea debido a la falla j, al tiempo t, en presencia del resto de las fallas, dado que no se presentó ninguna otra hasta ese tiempo[2].

$$h_j^*(t) = \lim_{\delta t \to 0^+} \frac{Pr\left(\{T \leq t + \delta t\} \bigcap \{C = j\} \,|T > t\right)}{\delta t}] \qquad (1)$$

Para $j = 1, 2, \ldots, p$ y $t > 0$.

Donde $Pr\left(\{T \leq t + \delta t\} \bigcap \{C = j\} \,|T > t\right)$, es la probabilidad de morir debido a la causa j, en $(t, t + \delta t)$. Todos los riesgos actúan en este intervalo porque se sobrevive a todas las causas posibles de falla al tiempo t.

2.2. Función de incidencia acumulada

El análisis de riesgos competitivos utiliza la función de incidencia acumulada para estudiar y determinar todos aquellos factores relacionados con la incidencia de un tipo de falla. En términos generales, es el número de casos nuevos de una enfermedad que se desarrolla en una población a lo largo de un periodo de tiempo determinado.

$$\text{IA} = \frac{\text{Número de casos nuevos a lo largo de un periodo}}{\text{Población en riesgo de enfermedad al inicio del periodo}}$$

La función de incidencia acumulada actúa en grupos de personas que están expuestas a diversas enfermedades.

[2]Lindqvist H. (2006), A review of Competing Risks, Department of Mathematical Sciences, Norwegian University of Science and Technology.

El riesgo es la probabilidad de que algún individuo del grupo contraiga una enfermedad durante el periodo de tiempo determinado. La incidencia acumulada es una medida o estimación del riesgo promedio. Se utiliza el promedio de la población para estimar el riesgo individual[3]. Para calcular la función de incidencia acumulada se necesita especificar un periodo de tiempo, porque la cantidad de sujetos expuestos a enfermedades varía con el paso de éste.

2.2.1. Estimación no paramétrica de la función de incidencia acumulada sin censura

Para el caso donde no hay censura, se utiliza una estimación empírica de la función de incidencia acumulada para el evento de interés del tipo j, la cual se puede expresar como:

$$F_j(t) = \frac{\text{Número de observaciones (T}\leq t, C = j)}{n}$$

Donde n, es el número total de observaciones.

Existe la posibilidad que entre las n observaciones algunos sujetos no experimenten alguna de las p fallas.

Si los tiempos son ordenados y sin censura $\{t_1 < t_2 < \ldots < t_r\}$. Se define d_{jk} como el número de eventos del tipo j, que ocurren al tiempo t_k. El número de sujetos que están bajo riesgos al tiempo t_k, se denota como n_k, y $S(t)$ es el estimador Kaplan-Meier (probabilidad de no presentar ningún evento al tiempo t).

La función de incidencia acumulada se puede calcular como la suma sobre todos los t_k, o sea, las probabilidades de observar el evento j, al tiempo t_k, estando bajo riesgo. Esto quiere decir, que no se experimentó ningún evento antes del tiempo t_k.

[3]Haesook, T. K. pag. 35, (2007).

La probabilidad de permanecer libre de fallas antes del tiempo t_k, se expresa como $S(t_{k-1})$. De lo anterior, se puede deducir que la probabilidad conjunta de estar libre de fallas inmediatamente antes del tiempo t_k, y experimentar un evento de tipo j, al tiempo t_k, se expresa como:

$$F_j(t) = \sum_{k,t_k \leq t} h^*_{jk} S(t_{k-1}) \qquad (3)$$

Donde h^*_{jk} es la función de riesgo de causa-específica para el evento j, al tiempo t_k. Si no se experimentó la falla al tiempo t_{k-1}, la probabilidad de que suceda un evento del tipo j, en el intervalo $(t_k - \delta)$ al tiempo t_k se estima como: $\left(\frac{d_{jk}}{n_k}\right)$.

$$F_j(t) = \sum_{j,t_j \leq t} \frac{d_{jk}}{n_k} S(t_{k-1}) \qquad (4)$$

El estimador de la función de incidencia acumulada para una falla j, no sólo depende del número de sujetos que lo han experimentado sino también de aquellos que no han tenido ningún tipo de evento[4]. Esta función representa la probabilidad de que un sujeto pueda experimentar una falla del tipo j, al tiempo t.

2.2.2. Estimación de la función de incidencia acumulada con censura

Un estimador para h^*_{jk} y la incidencia acumulada, se deriva de la sección anterior. En el intervalo $[t_{k-1}, t_k)$ puede haber m_k, observaciones censuradas en los tiempos $\{t_{k_1} < t_{k_2} < \ldots < t_{k_m}\}$.

La función de verosimilitud se puede escribir como:

$$L = \left[\prod_{k=1}^{r}\prod_{j=1}^{p} \left\{F_j(t_k) - F_j\left(t_{k-1}\right)\right\}^{d_{jk}}\right] \left[\prod_{k=1}^{r+1}\prod_{l=1}^{m_k} S(t_{kl})\right] \qquad (5)$$

[4]Competing Risk a practical perspective. pag. 55, (2006).

En términos de la función de riesgo de causa-específica se puede expresar como:

$$L = \prod_{k=1}^{r} \prod_{j=1}^{p} h_{jk}^{d_{jk}} \left(1 - h_k\right)^{n_k - d_k} \qquad (6)$$

Donde h_k, es el riesgo de cualquier tipo de evento al tiempo t_k.

Es común que la log-verosimilitud se obtenga con respecto al riesgo de la función de causa-específica, h_{jk}^*,

$$h_{jk}^* = \frac{d_{jk}(1 - h_k)}{n_k - d_k} \qquad (7)$$

donde:

$$h_k = \sum_{j=1}^{p} h_{jk}^* \qquad d_k = \sum_{j=1}^{p} d_{jk}$$

Al sustituir h_k y d_k en (7), se obtiene:

$$h_k = \frac{d_k(1 - h_k)}{n_k - d_k} \qquad (8)$$

El estimador para h_k, es $\widehat{h}_k = \frac{d_k}{n_k}$. Al sustituirlo en (7), se obtiene que $\widehat{h}_{jk} = \frac{d_{jk}}{n_k}$. El nuevo estimador se puede sustituir en (3) y se genera el estimador de la función de incidencia acumulada:

$$\widehat{F_j}(t) = \sum_{t_k \leq t} \frac{d_{jk}}{n_k} \widehat{S}(t_{k-1}) \qquad (9)$$

Donde $\widehat{S}(t)$ es el estimador de Kaplan-Meier, éste se hubiera obtenido tomando en cuenta que todos los eventos son del mismo tipo. La estimación

de máxima verosimilitud de la función de incidencia acumulada es igual a la obtenida a través de la estimación de esta misma función[5].

2.3. Función de incidencia acumulada de causa-específica

La función de incidencia acumulada de causa-específica está determinada por la distribución conjunta del par (T, C). Retomando la ecuación (1), la probabilidad condicional de experimentar una falla debido a la causa j, en el intervalo $(t, t+\delta t)$, condicionado a que se ha sobrevivido (libre de cualquier evento) hasta el tiempo t, y en presencia de todas las causas actuando simultáneamente, es aproximadamente: $h_j(t)\delta t$,

$$Pr\left(\{T < t + \delta t\} \bigcap \{C = j\} \,|T > t\right) \approx h_j(t)\delta t \qquad (10)$$

La probabilidad no condicional de morir debido a la falla j, en el intervalo $(t, t + \delta t)$ es, $h_j(t)S(t)dt$. Esto define a la función de incidencia acumulada de causa-específica al tiempo t, ocasionada por la falla j, en presencia de todas las causas (actuando simultáneamente):

$$F_j(t) = Pr\left(\{T \leq t\} \bigcap \{C = j\}\right) = \int_0^t h_j(u)S_T(u)du \quad \text{con} \quad t > 0 \qquad (11)$$

para $j = 1, 2, 3, \ldots, p$.

$S(t)$, es la función de supervivencia total y representa la probabilidad de sobrevivir a cualquier tipo de falla hasta el tiempo, t. Toma en cuenta p tipos de fallas. En el análisis de supervivencia únicamente considera un único tipo de falla.

[5]Competing Risk a practical perspective. pag. 58, (2006).

Análogamente, se define la función de supervivencia de causa-específica, como:

$$S_e(t) = Pr\left(\{T > t\} \bigcap \{C = j\}\right) = \int_t^{\infty} h_j(u)S(u)du \qquad (12)$$

Esta función describe la probabilidad de morir debido a la falla j, en un tiempo mayor a t.

Tsiatis[6] demostró que para cualquier función conjunta de tiempos de falla latentes:

$$S_{1,\ldots,p}\left(y_1, y_2, \ldots, y_p\right) = Pr\left(\bigcap_{j=1}^{p} \{Y_j > y_j\}\right) \qquad (13)$$

se tiene:

$$\frac{dS_e(t)}{dt} = \frac{\partial S_{1,\ldots,p}(y_1, \ldots, y_p)}{\partial y_j}|_{y_1=\cdots=y_p=t} \qquad (14)$$

Este resultado indica que cualquier función de decrementos múltiples: $S_{1,2,\ldots,p}(y_1, y_2, \ldots, y_p)$ determina de manera única al conjunto de funciones de supervivencia de causa-específica $\{S_e(t) : j = 1, 2, \ldots, p\}$.

La función de densidad de probabilidad de causa-específica se puede asociar con la de supervivencia $S_j(t)$, cuando ésta existe:

$$f_j(t) = \frac{-\partial S_{1,\ldots,p}(y_1, \ldots, y_p)}{\partial y_j}|_{y_1=\cdots=y_p=t} \qquad (15)$$

Esta función representa el riesgo incondicional de que se experimente una falla j al tiempo t. Derivándose la siguiente relación,

[6]Tsiatis, A. (1975).
A nonidentifiability aspect of the problem of competing risks, pag. 25.

$$h_j(t) = -\frac{\partial \log S_{1,\ldots,p}(y_1,\ldots,y_p)}{\partial y_j}|_{y_1=\cdots=y_p=t} = \frac{f_j(t)}{S(t)} \qquad (16)$$

Al despejar $f_j(t)$, de (16) y tomando la suma sobre $j = 1, 2, \ldots, p$. Se obtiene la función de densidad de probabilidad del tiempo de falla T,

$$\begin{aligned} \sum_{j=1}^{p} f_j(t) &= S(t)\sum_{j=1}^{p} h_j(t) \\ &= S(t)h(t) \\ &= f(t) \\ &= -\frac{dS(t)}{dt} \qquad (17) \end{aligned}$$

La distribución marginal de T, se obtiene con la función de distribución acumulada:

$$\begin{aligned} \sum_{j=1}^{p} F_j(t) &= \sum_{j=1}^{p}\int_0^t h_j(u)S(u)du \\ &= \int_0^t \sum_{j=1}^{p} h_j(u)S(u)du \\ &= \int_0^t h(u)S(u)du \\ &= -\int_0^t dS(u) \\ &= 1 - S(t) \\ &= F(t) \qquad (18) \end{aligned}$$

donde:

$$\begin{aligned} S(t) &= Pr(T > t) \\ &= \sum_{j=1}^{p} S_e(t) \qquad (19) \end{aligned}$$

S(t), es la función de distribución de supervivencia marginal para T. La proporción de fallas debido a la causa C_j, denotada como $\tau_j(j)$ en $(j \in C)$, se obtiene de la distribución marginal de C.

$$\begin{aligned} \tau_j &= Pr\,(C = j) \\ &= Pr\left(\{T < \infty\} \bigcap \{C = j\}\right) \\ &= \int_0^\infty h_j(u) S(u) du \\ &= F_j(\infty) = S_e(0) \qquad (20) \end{aligned}$$

$$\text{con} \quad \tau_j > 0 \quad \text{y} \quad \sum_{j=1}^{p} \tau_j = 1$$

Se puede deducir que existe relación entre la función de incidencia acumulada de causa-específica y la de supervivencia de causa-específica:

$$F_j(t) + S_e(t) = \tau_j \qquad \text{para} \quad j = 1, 2, \ldots, p$$

De la ecuación (18) se obtiene:

$$F(t) + S(t) = 1 \qquad (21)$$

En la ecuación (21) se puede observar que las funciones: $F_j(t)$ y $S_e(t)$ no son distribuciones propias, porque cada τ_j es menor a uno. Sin embargo, se pueden definir distribuciones condicionales en términos de las funciones de causa-específica. Por ejemplo, la función de distribución de supervivencia condicional (en presencia de todas las causas) asociada a la falla j, y denotada como S_j^*(t):

$$\begin{aligned} S_j^*(t) &= Pr\,(T > t | C = j) = \frac{S_e(t)}{S_e(0)} = \frac{1}{\tau_j} S_e(t) \\ &= \frac{1}{\tau_j} \int_t^\infty h_j(u) S(u) du \quad \text{con} \quad j = 1, 2, \ldots, p \qquad (22) \end{aligned}$$

entonces:

$$F_j^*(t) = 1 - S_j^*(t) = Pr\left(T \leq t | C = j)\right] \qquad (23)$$

$F_j^*(t)$ expresa la probabilidad condicional de presentar una falla antes del tiempo t, en presencia de todas.

Para obtener la probabilidad de tener una falla debido a la causa j, después del tiempo t, dado que se ha sobrevivido a todas las causas de falla posibles hasta ese momento, la función correspondiente de distribución de probabilidad de la ecuación número (13) es:

$$\begin{aligned} f_j^*(t) &= -\frac{dS_j^*(t)}{dt} \\ &= -\frac{1}{\tau_j}\left(\frac{dS_e(t)}{dt}\right) \\ &= \frac{1}{\tau_j} h_j(t) S(t) \qquad (24) \end{aligned}$$

por ello,

$$h_j(t) = -\frac{1}{S(t)}\frac{dS_e(t)}{dt} = -\frac{dS_e(t)/dt}{\sum_{k=1}^{p} S_k(t)} \qquad (25)$$

Por otro lado, la función de riesgo de S_j^* es:

$$\begin{aligned} h_j^*(t) &= \lim_{\delta t \to 0^+} \frac{1}{\delta t} Pr\left(t < T + \delta t | T > t, C = j\right) \\ &= -\frac{d \log S_j^*(t)}{dt} = \frac{f_j^*(t)}{S_j^*(t)} \\ &= \frac{h_j(t) S(t)}{\int_t^\infty h_j(u) S(u) du} \\ &= -\frac{1}{S_e(t)}\frac{dS_e(t)}{dt} \qquad (26) \end{aligned}$$

A partir de las ecuaciones (25) y (26), se obtiene:

$$\frac{h_j(t)}{h_j^*(t)} = \frac{S_e(t)}{S(t)} = \frac{S_e(t)}{\sum_{j=1}^{p} S_e(t)} = Pr\left(C = j | T > t\right) = \tau_j(t) \qquad (27)$$

Esta igualdad representa la probabilidad condicional de que la causa j ocurra, después del tiempo t, dado que se ha sobrevivido a todas las causas hasta ese tiempo.

Capítulo 3

Modelos para estimar riesgos competitivos

Introducción

El objetivo de este capítulo es mostrar algunos de los modelos que se utilizan para estimar riesgos competitivos. En la primera sección se desarrolla el primer modelo que se utilizó en el análisis de riesgos competitivos, el complemento del estimador Kaplan-Meier. La utilización de dicho modelo parte del análisis de supervivencia y la modificación de uno de los métodos más importantes y utilizados en dicho análisis.

En la segunda sección se muestra el método de Gray, el cual es un modelo que estima riesgos competitivos a través de pruebas de covariables. El autor de dicho modelo introdujo la teoría de r-muestras y construyó un paquete estadístico que hace estimaciones a través del programa `r.project`.

El método de Gray es el modelo utilizado en este trabajo para estimar datos y elaborar una simulación que muestran la aplicación de las funciones del paquete `cmprsk` del programa `r.project`.

3.1. Complemento del estimador Kaplan-Meier

En los primeros trabajos que se llevaron a cabo bajo el enfoque de riesgos competitivos se utilizó el complemento de Kaplan-Meier como método de análisis.

$$\begin{aligned} \hat{P}(t) &= 1 - \hat{S}(t) \\ &= 1 - \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right) \end{aligned}$$

Se ha observado que los resultados analizados bajo el enfoque del complemento de Kaplan-Meier $(1-KM)$ están sesgados. Comparando el método de incidencia acumulada con el complemento de Kaplan-Meier, se mostró que la probabilidad con $1 - KM$, es mayor que la obtenida con la función de incidencia acumulada.

Utilizar el complemento del estimador de Kaplan-Meier en presencia de riesgos competitivos, no es la mejor opción, porque la estimación obtenida no cumple con la característica que define a un evento de riesgo competitivo, porque sólo toma en cuenta un único evento de interés en la estimación. Analizar un evento en ausencia de los demás genera probabilidades más altas.

El *Dapartment of Statistics of the University of Toronto*[1] en conjunto con el *Hospital Princess Margaret de Canadá*, hicieron un estudio, donde se muestra que al estimar datos con el complemento del estimador de Kaplan-Meier (1-KM), se obtiene una probabilidad mayor en comparación con la estimación de la función de incidencia acumulada.

El estudio se realizó a 30 pacientes que ingresaron al hospital por varicela, se registró el tiempo en días para ser dados de alta del hospital. El periodo de tiempo que se consideró razonable para la alta hospitalaria fue de 1 a 28 días como máximo.

[1]Competing Risks: A Practical Perspective M. Pintilie. (2006).

El evento de interés en este estudio ha sido que el paciente sea dado de alta dentro del periodo establecido de 1 a 28 días. El riesgo competitivo, es que ocurra la muerte antes de la alta hospitalaria causada por alguna de las p fallas. Los 30 pacientes fueron observados hasta el momento en que ocurrió el evento de interés o el riesgo competitivo (muerte por diversas fallas). No hubo observaciones censuradas. El estudio tiene como fin encontrar la probabilidad conjunta que se produce al presentarse el evento de interés (alta hospitalaria a lo más en 28 días).

La tabla 1 muestra la información ordenada de los pacientes: número de identificación, días internados y estatus. El número 1, indica que se observó el evento de interés y el número 2 se refiere a la muerte del paciente.

El primer análisis de los datos se hizo bajo el enfoque de la función de incidencia acumulada, cabe resaltar que el orden de los datos no afecta el cálculo de esta función:

$$F_1(28) = Pr(T \leq 28, C = 1)$$

- T= tiempo del primer evento (alta hospitalaria o muerte).
- C= tipo de evento.

C, tiene dos significados: alta hospitalaria (1) o muerte (2).

$$\begin{aligned} \widehat{F_1}(28) &= \frac{\text{Número de pacientes dados de alta a lo más en 28 días}}{\text{Número total de pacientes}} \\ \widehat{F_1}(28) &= \frac{13}{30} = .333 \end{aligned}$$

El resultado indica que la probabilidad de ser dado de alta del hospital a lo más el día 28, es igual a .333, tomando en cuenta el riesgo de morir durante ese periodo debido a cualquier otra causa.

Individuo	Días	Tipo de falla
1	2	1
2	1	2
3	3	1
4	4	1
5	5	1
6	8	1
7	9	1
8	10	2
9	11	2
10	13	1
11	14	2
12	15	1
13	17	2
14	19	1
15	21	1
16	23	1
17	24	1
18	25	2
19	28	1
20	29	1
21	32	1
22	34	1
23	35	1
24	36	2
25	37	2
26	38	2
27	39	1
28	48	1
29	50	1
30	63	2

Tabla 1: Pacientes internados por varicela.

El segundo enfoque utilizado en este estudio, para analizar los datos, es el complemento del estimador de Kaplan-Meier. Bajo esta función, lo primero que se debe hacer, es construir el estimador común KM, para posteriormente realizar la resta $(1-KM)$. El estimador de Kaplan-Meier se puede obtener a través del paquete `Surv` y la función `survfit` del programa `r.project`.

Primero se debe construir una variable que guarde al estimador de KM. La función que lo obtiene es `Surv`, la cual tiene dos parámetros `T= verdadero` y `S==1`, este último le dice a la función que el evento de interés es igual a 1. Anteriormente se definió que el número 1 representa la alta hospitalaria a lo más en 28 días. Poner `T` dentro de los parámetros sirve para hacer verdadera la afirmación del segundo. La condición construida con `Surv` se debe aplicar a todos lo datos que se van a estudiar, esto se hace agregando `.all` a la variable construida al principio y se le aplica la función `survfit`, para observar todos los valores que guarda `survfit`, se utiliza la función `summary()`. Finalmente para obtener el histograma de los datos analizados y observar los resultados gráficamente se utiliza la función `plot()`.

```
gente<- Surv(T,S==1)
gente.all <-survfit(gente~1, conf.type=''plain'',
conf.int=.9)
gente.all
summary(gente.all)
plot(gente.all)
```

La probabilidad de supervivencia (alta hospitalaria a los más en 28 días) es de `.5953`. Una vez obtenido el estimador Kaplan-Meier se aplica la fórmula (1-KM).

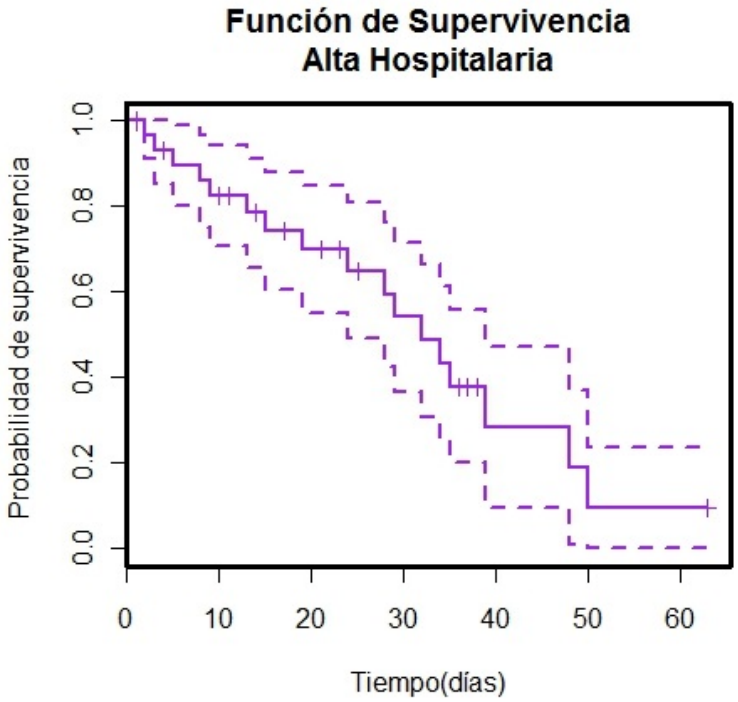
$$\begin{aligned} \text{Completemento de Kaplan-Meier} &= 1 - KM \\ &= 1 - .5953 \\ &= .4047 \end{aligned}$$

Existe una diferencia considerable en el valor de la probabilidad entre ambos métodos: función de incidencia acumulada y complemento del estimador de Kaplan-Meier.

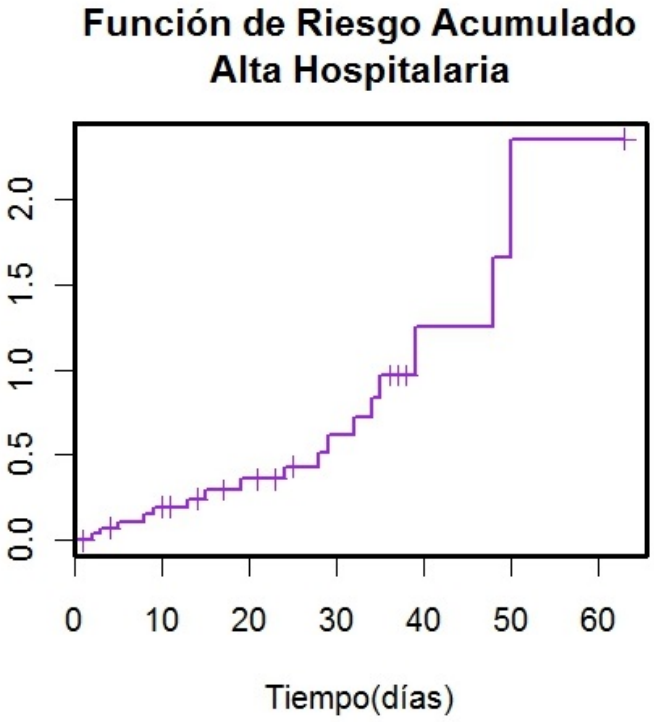
- Función de incidencia acumulada, $FIA =$ `.333`
- Complemento del estimador de Kaplan-Meier, $(1 - KM) =$ `.4047`

Esta diferencia se origina porque al estimar el complemento de Kaplan-Meier, se ignora todo evento distinto al de interés, los demás riesgos se agrupan en uno solo y no se toman en cuenta. Este resultado se puede interpretar como la probabilidad de un evento que va más allá del tiempo, t, pero no es un resultado que se pueda utilizar para obtener alguna conclusión sobre el estudio.

La siguiente gráfica es el histograma del estimador Kaplan-Meier que se obtiene con la función `plot(gente.all)` y corresponde a la función de supervivencia.



Dentro del parquete `Surv`, existe una función que se llama `cumhaz`, la cual permite obtener el riesgo acumulado en cada punto del análisis, de igual forma se puede obtener su histograma con la función `plot()`.



Censurar a los pacientes que presentan la falla a partir de un riesgo competitivo no es una opción adecuada, porque la estimación de la probabilidad de falla es lo que se está buscando. Esto supone implícitamente que el evento de interés todavía es posible más allá del tiempo en que la censura se produjo. Si un paciente presenta la falla a partir de un riesgo competitivo, el evento de interés ya no es posible y la contribución potencial de la estimación de este paciente es cero.

Los métodos de estimación para estudios con riesgos competitivos se han ido desarrollando, la estimación bajo la función de incidencia acumulada introducida por Fine JP, Gray RJ[2], se deriva de una modificación al modelo de Cox, para tomar en cuenta los riesgos competitivos, la cual ayudó en la construcción del paquete `cmprsk` del programa `r.project`.

[2]A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc. pag. 45, (1999).

3.2. Método de Gray (prueba de covariables)

Gray R[3] introdujo la teoría de r-muestras, la cual se deriva del modelo de riesgos proporcionales de Cox[4] (comenzando con el caso de no censuras) se presenta el caso general para datos incompletos. Mostrando lo que usualmente sucede en los análisis: un porcentaje de los sujetos presentan el evento de interés, otro se registro bajo un riesgo competitivo y el resto son censuras. Definió el siguiente modelo:

$$\gamma(t, x) = \gamma_0(t) \exp^{\beta x} \qquad (28)$$

Donde,

- γ. Representa el riesgo de la función de incidencia acumulada.
- γ_0. Es el riesgo inicial de la función de incidencia acumulada.
- x. Es un vector de covariables.
- β. Es un vector de coeficientes.

Partiendo de que in individuo i con un determinado conjunto de covariables $x_{1_i}, x_{2_i}, ..., x_{m_i}$. Se puede descomponer en dos partes: una que implica al tiempo, pero no a las covariables, y otro que implica a las covariables pero no al tiempo.

$$h_j = exp\{\beta_j x_{1_j} + ...\beta_m x_{m_j}\} h_0(t)$$

Donde h_0 es el riesgo inicial y $\beta_1, \beta_2, ..., \beta_m$ son los parámetros de los coeficientes que deben ser estimados. Antes de $t_1, t_2, ..., t_r$ los tiempos de falla únicos y ordenados. Cuando no existen vínculos entre las estimaciones de los coeficientes $\beta_1, \beta_2, ..., \beta_m$ se obtienen a través de la máxima verosimilitud:

[3] Analyzing Survival Data with Competing Risks, (2012).

[4] Para más información revisar Apéndice A. Estimador de Kaplan-Meier y Riesgos Proporcionales de Cox.

$$L(\beta_1, \beta_2, ..., \beta_m) = \prod_{k=1}^{r} \frac{exp\{\beta_1 x_{1_k} + ...\beta_m x_{m_k}\}}{\sum_{j \in R_k} exp\{\beta_1 x_{1_j} + ...\beta_m x_{m_j}\}}$$

Donde el producto se toma sobre todos los tiempos de falla t_k y R_k representa el conjunto de individuos que aún están bajo riesgo al tiempo t_k.

Partiendo de la función anterior, la verosimilitud parcial es similar a la utilizada en el modelo de riesgos proporcionales de Cox, para una covariable x y está dada por:

$$L(\beta) = \prod_{k=1}^{r} \frac{\exp(\beta x_k)}{\sum\limits_{j \in R_k} w_{kj} \exp(\beta x_j)} \qquad (29)$$

El producto asume todos los r puntos de tiempo ($t_1 < t_2 < \ldots < t_r$) donde el evento de interés es observado.

$$L(\beta_1, \beta_2, \ldots, \beta_m) = \prod_{k=1}^{r} \left(\frac{\exp\{\beta_1 x_{1k} + \ldots + \beta_m x_{mk}\}}{\sum\limits_{j \in R_k} \exp\{\beta_1 x_{1j} + \ldots + \beta_m x_{mj}\}} \right) \qquad (30)$$

$$L(\beta) = \prod_{k=1}^{r} \frac{\exp(\beta x_k)}{\sum\limits_{j \in R_k} w_{kj} \exp(\beta x_j)} \qquad (31)$$

Lo que diferencia la ecuación (30) de la (31), es que el conjunto de riesgo R_k, se define de forma diferente. En la ecuación (31), se añade (w_{kj}), y representa al conjunto de riesgo formado por quienes no han presentado el evento de interés al tiempo t, y por aquellos que han presentado el evento de interés debido a un riesgo competitivo al tiempo t.

$R_k = \{j; T_j \geq t \quad \text{o} \quad (T_j \leq t \quad \text{el sujeto experimentó un riesgo competitivo})\}$

Quienes experimentaron otro tipo de evento permanecen en el conjunto de riesgo al tiempo t_r, y w_{kj} se define como:

$$w_{kj} = \frac{\widehat{G}(t_k)}{\widehat{G}(\text{mín}(t_k, t_j))} \qquad (32)$$

Donde $\widehat{G}$ es el estimador de Kaplan-Meier de la función de supervivencia que contiene censuras. Esta distribución se define como (T_j, C_j). Donde T_j es el tiempo del primer evento y C_j es igual a 1, si no se observó ningún evento e igual a 0, si se observa cualquier tipo de evento.

Para cada tiempo donde se presente el evento de interés se le anexará el subíndice k, mientras que al conjunto de personas en situación de riesgo se indetificará con el subíndice j. Este conjunto incluye a todos aquellos que no presentan ningún tipo de falla al tiempo t_k, así como a los que presenten un evento por riesgo competitivo antes del tiempo t_k. El valor es igual a 1, para el primero y menor o igual a 1, para el segundo. Quienes experimenten un evento de riesgo competitivo no participan plenamente en la verosimilitud parcial.

Al derivar la ecuación (31), se obtiene la estadística de puntajes:

$$U(\beta) = \sum_{k=1}^{r} \left\{ x_k - \frac{\sum_{j \in R_k} w_{kj} x_j \exp(x_j \beta)}{\sum_{j \in R_k} w_{kj} \exp(x_j \beta)} \right\} \qquad (33)$$

El estimador de β, denotado como $\widehat{\beta}$, es el valor máximo de la función $U(\beta)$. Por lo tanto, la ecuación (33) puede escribirse para un conjunto de covariables y se pueden incluir más. Por lo tanto, se puede trabajar con p fallas en el caso de riesgos competitivos.

Capítulo 4

Riesgos competitivos con R

Introducción

En la primera sección de este capítulo se muestran las funciones que contiene el paquete `cmprsk` del programa `r.project`, así como las variables que las componen. La paqueteria `cmprsk` fue desarrollada por el Dr. Robert Gray[1].

En la segunda sección se desarrolla todo el procedimiento que se utilizó para analizar los datos `ccup.csv`, con el fin de explicar cómo se utilizan las funciones descritas en la primera sección de este capítulo. La base de datos estudiada en este trabajo, contiene información sobre mujeress que padecieron cáncer de cuello uterino primario, y fue elaborada por el *Department of Statistics of the University of Toronto* en conjunto con el Hospital *Princess Margaret de Canadá*.

[1]Ph.D. Professor of Biostatistics, Department of Biostatistics. University of Toronto.

4.1. Función de incidencia acumulada con el paquete `cmprsk`.

La función de incidencia acumulada (FIA) es una herramienta que sirve para analizar datos de riesgos competitivos, es uno de los diferentes métodos que son utilizados para estudiar este tipo de datos. En los capítulos anteriores se describió la teoría que sustenta este enfoque. A continuación se describen las funciones que contiene la libreria `cmprsk` del programa `r.project`, la cual sirve para analizar datos de riesgos competitivos bajo el enfoque de la función de incidencia acumulada.

4.1.1. Función `crr`

Se utiliza para estudiar la regresión basada en las funciones de causa-específica de los modelos en riesgos competitivos.

Sintaxis:

```
crr(ftime, fstatus, cov1, failcode=1, subset)
```

Argumentos empleados en la sintaxis:

- **ftime:** es un vector de tiempos de falla o censuras.
- **fstatus:** es un vector con código único para cada tipo de error y un código separado para las observaciones que son censura.
- **cov1:** es una matriz (Nobs x ncovs) de las covariables fijas (ya sea `cov1`, `cov2`, o ambas). Estas covariables también aparecerán en cov1 para proporcionar un efecto de apoyo en los riesgos además de un tipo de interacción.
- **failcode:** es un código de `fstatus` que indica el tipo de falla (evento de interés).
- **subset:** un vector lógico que especifica un subconjunto de casos para incluir en el análisis.

Se adapta el modelo de regresión de riesgos de subdistribución proporcionales. Este modelo evalúa directamente el efecto de las covariables sobre la causa-específica de un tipo particular de falla en un entorno de riesgos competitivos. El método aplicado aquí se describe como la ecuación ponderada de estimación.

4.1.2. Función `print.crr`

Esta función obtiene el histograma para `crr`.

Sintaxis:

```
print(x, ...)
```

Argumentos empleados en la sintaxis:

- **x:** objeto `crr` (salida de `crr()`)

Muestra el estado de convergencia, los coeficientes estimados, los errores estándar estimados, y los p-values para la prueba de hipótesis de que los coeficientes individuales sean iguales a 0.

4.1.3. Función `predict.crr`

Esta función es un método de predicción para `crr`. Estima las funciones de causa-específica de la salida de `crr`.

Sintaxis:

```
predict(object, cov1, cov2, ...)
```

Argumentos empleados en la sintaxis:

- **object:** es la salida de `crr`.

- **cov1, cov2:** cada fila de cov1 y cov2 son un conjunto de valores (covariables) donde la causa-específica debe ser estimada. Las columnas de cov1 y cov2 deben estar en el mismo orden como en la original `crr`. Cada uno debe darse si está presente en la función `crr`.

Calcula $\{1 - e^{-B(t)}\}$, donde B(t) es el riesgo de la causa-específica acumulada estimada que se obtiene para los valores de las covariables específicas a partir de la estimación de tipo Breslow del riesgo base y los coeficientes de regresión estimados.

4.1.4. Función `plot.predict.crr`

Esta función describe la función `predict.crr.` a través de un histograma.

Sintaxis:

```
plot(x, lty=1:(ncol(x)-1), color=1, ylim=c(0, max(x[, -1])),
xmin=0, xmax=max(x[, 1]), ...)
```

Argumentos empleados en la sintaxis:

- **x:** es la salida de `predict.crr`.

Dibuja las funciones de causa-específica estimadas por `predict.crr` en un histograma.

4.1.5. Función `cuminc`

Esta función estima las funciones de incidencia acumulada de los datos que están en riesgos competitivos y prueba la igualdad entre todos los grupos.

Sintaxis:

```
cuminc(ftime,fstatus,subset)
```

Argumentos empleados en la sintaxis:

- **ftime:** es una variable que representa el tiempo de falla.
- **fstatus:** es una variable con códigos distintos para diferentes tipos de falla y para observaciones censuradas.
- **subset:** es un vector lógico que específica un subconjunto de casos para incluir en el análisis.

Es una lista con los componentes (estimaciones) de la causa-específica para cada falla en cada grupo, y un componente "tests" que da el estadístico y p-value para comparar la causa-específica de cada falla entre los grupos (si el número de grupos es mayor a uno). Los componentes que dan las estimaciones tienen nombres que son una combinación del nombre del grupo y la falla.

4.1.6. Función `plot.cuminc`

Es una función que describe la función `cuminc` a tráves de un histograma.

Sintaxis:

```
plot(x, main=, curvlab, ylim=c(0, 1), xlim, wh=2,xlab="Years",
ylab="Probability", lty=1:length(x), color=1,
lwd=par('lwd'),...)
```

Argumentos empleados en la sintaxis:

- **x:** es una lista, con cada componente que representa una curva en el histograma. Cada elemento de x en sí es una lista cuyo primer factor proporciona los valores de x y el segundo representa los de y.
- **main:** es el título del histograma.
- **curvlab:** son las etiquetas de las curvas.

4.2. Aplicación de las funciones del paquete `cmprsk` a la base de datos `ccup.csv`

El objetivo de esta sección es mostrar como se utilizan algunas de las funciones que componen al paquete `cmprsk`. Se utiliza una base de datos llamada (`ccup.csv`), con la cual se muestra la construcción de las funciones para analizar un caso de riesgos competitivos.

La base **ccup_rc.csv** se construyó entre los años $1994 - 2000$, cuenta con 109 individuos. Su edad oscila entre los 23 y 78 años, el sexo de todos los sujetos es femenino y la información recabada se distribuye en trece variables. Como en todo análisis, es muy importante conocer con qué se está trabajando. La enfermedad que se encuentra bajo estudio es el cáncer de cuello uterino primario.

El cérvix o cuello uterino es la parte inferior del útero que forma el canal que lleva a la vagina. La mucosa que recubre el cérvix está en continuidad con la vagina y se denomina ectocérvix, mientras que la que recubre el conducto o canal cervical que lleva hasta la cavidad del cuello uterino, se denomina endocérvix. En este tipo de cáncer, la mayor parte de los tumores surgen en la zona donde se une el ectrocérvix con el endocérvix dando lugar a carcinomas de células escamosas. El cáncer se produce cuando las células normales del cuello del útero empiezan a transformarse y crecen de manera descontrolada.

El cáncer de cuello uterino es el segundo tipo de cáncer más frecuente en la mujer, y prácticamente todos los casos están relacionados con la infección del Virus del Papiloma Humano (VPH). En México, en 2013 ocurrieron 269 mil 332 defunciones en mujeres, de las cuáles los tumores malignos representan el 13.8 % (37 mil 361) de esas muertes. Dentro de las neoplasias con mayor número de defunciones en mujeres, el cáncer de mama y el cuello uterino ocasionaron en conjunto el 25 % de todas las defunciones por cáncer en mujeres. Es decir, 1 de cada 10 muertes por cáncer en mujeres mexicanas es debida a cáncer de cuello uterino[2].

[2]Instituto Nacional de Cancerología.

Variables contenidas en la base de datos:

1. **edad** = Edad (años).
2. **hgb** = Hemoglobina(g/l).
3. **tamtum** = Tamaño del tumor (cm).
4. **IFP** = Presión del fluido intersticial (marcador, mmHg).
5. **HP5** = Marcador de hipoxia (porcentaje de las mediciones de menos de 5 mm de Hg).
6. **resp** = Respuesta después del tratamiento: CR = respuesta completa, NR = No respondió.
7. **pelrec** = Enfermedad pélvica observada: Y=Sí, N=No.
8. **disrec** = Enfermedad distante observada: Y=Sí, N=No.
9. **survtime** = Tiempo desde el diagnóstico hasta la muerte o el último de seguimiento.
10. **stat** = Estado al último seguimiento: 0 = Vivo, 1 = Muerto.
11. **dftime** = Tiempo a partir del diagnóstico hasta la primera falla (sin respuesta al tratamiento, recaída o muerte) o el último seguimiento.
12. **dfcens** = Variables censuradas: 1 = Falla, 0 = Censurado.
13. **pelvi** = Nodo pélvico implícado: N = Negativo/Equivoco, Y = Positivo.

En el análisis de riesgos competitivos existe más de un evento de interés, en estos casos se pueden transformar en covariables algunas variables que están relacionadas con la de interés, para determinar si éstas influyen o no en el evento que se está estudiando.

En este estudio, las variables de resultado son: respuesta al tratamiento, recaída y muerte. Para definir el primer evento, respuesta al tratamiento, se construyó una variable que indica si el paciente con cáncer de cuello uterino

primario respondió o no al tratamiento (quimioterapia más radioterapia). La variable se definió como `resp` y solamente tiene dos posibles resultados: el paciente tuvo respuesta completa al tratamiento `resp=CR` o el paciente no respondió al tratamiento `resp=NR`.

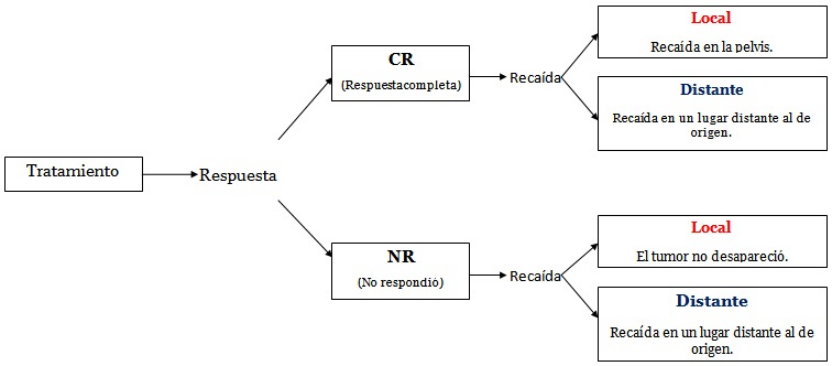


Diagrama 4.1: Eventos.

El diagrama 4.1 muestra que la recaída local tiene dos significados: recaída en la pélvis o el tumor no desapreció. El resultado depende de si el paciente respondió o no al tratamiento inicial. La variable recaída local `(pelrec)` es un riesgo competitivo. Se puede comparar dicha variable con recaída distante `(disrec)`, la cual es una variable que significa lo mismo, tanto en el caso cuando el paciente tiene una respuesta completa al tratamiento, como cuando no la tiene. A diferencia de la variable recaída local que tiene más de un significado.

La variable `resp` ayuda a definir el estado de dos variables: `disrec` y `pelrec`. La primera significa recaída distante y la segunda recaída local. Cuando `resp` es igual a `NR`, el tumor no desapareció, entonces `pelrec=Y`, y el cáncer pudo o no haber reincidido en un lugar distante: `disrec=Y` o `disrec=N`.

Si un paciente tuvo respuesta completa al tratamiento `CR`, la enfermedad pudo haber surgido de nuevo, esto quiere decir, que el cáncer no tardó en reincidir en un lugar cercano al de origen, como la pelvis, entonces `pelrec=Y`, con o sin recaída distante: `disrec=Y` o `direc=N`.

A continuación se muestran los casos posibles de `pelrec` y `disrec`, definidos por el resultado de la variable `resp` (respuesta completa `CR` o no respondió `NR`):

`resp=CR`

1. `pelrec=Y disrec=Y`
2. `pelrec=Y disrec=N`
3. `pelrec=N disrec=Y`
4. `pelrec=N disrec=N`

Cuando `resp` es igual a `CR`, se tienen cuatro posibles escenarios, de los cuales sólo se tomán en cuenta en este análisis los dos primeros.

El tercer caso se omite porque anula al riesgo competitivo, debido a que `pelrec` es igual a `N` (sin recaída local) y la variable `disrec`, que representa recaída distante, tiene el mismo significado tanto para `resp=CR` como para `resp=NR`.

El cuarto caso se elimina porque significa que el paciente está libre de enfermedad después de recibir el tratamiento y no presenta ningún tipo de recaída. Lo cual, no es tema de estudio de este trabajo. Por lo tanto, los escenarios posibles que quedan para `resp=CR` son:

`resp=CR`

1. `pelrec=Y disrec=Y`
2. `pelrec=Y disrec=N`

Cuando `resp=NR`, quiere decir, que el tratamiento no funcionó en la mujer, lo que significa que nunca estuvo libre de enfermedad, después de recibir el tratamiento inicial. Por lo tanto, existen únicamente dos escenarios, recaída local con recaída distante y recaída local sin recaída distante.

La variable `pelrec` en el caso de `resp=NR`, nunca puede tener valor igual a `N`, porque el mismo resultado `NR`, indica que el paciente nunca estuvo libre

de la enfermedad y por lo tanto, no respondió al tratamiento.

Entonces, cuando `pelrec=Y`, puede significar dos cosas: la primera es que el paciente no respondió al tratamiento o la enfermedad reincidió en un lugar no lejano al de origen, en la pelvis. La variable `resp` indica de qué caso se trata.

Para este conjunto de datos la muerte se produjo antes de una recaída, por lo tanto, no es un evento de interés. El tiempo del evento está cálculado en años, comenzando en la fecha de diagnóstico hasta que se presenta la primera falla.

La enfermedad estudiada es el cáncer de cuello uterino en su fase primaria, esto implica que los pacientes desarrollaron un tumor en la parte más baja del útero. Para realizar el estudio se analizó la influencia de las siguientes variables relacionadas con el tumor: hipoxia (`HP5`) y presión del fluido intersticial `IFP`.

- La hipoxia se refiere a la disminución en la cantidad de oxígeno suministrado por la sangre a los órganos.
- El líquido intersticial o líquido del intersticio, sirve para rellenar la parte vacía entre las células y los capilares sanguíneos.

El nivel de oxigenación se grabó en milímetros de mercurio (mmHg) y se midió en cada tumor de 25 a 30 veces a lo largo del estadio, con 3 ó 4 estadios por tumor. El marcador de hipoxia fue definido como el porcentaje de mediciones en un tumor que tenía nivel de oxígeno inferior a 5 mmHg. La IFP se midió en diversos lugares del tumor y se cálculo para obtener un valor medio por paciente.

La hipótesis con la que parte este análisis es que las variables: `HP5` e `IFP`, influyen en la recaída local de un paciente con cáncer de cuello uterino primario. Para rechazar o aceptar esta hipótesis se utilizan las dos principales funciones del paquete `cmprsk`: `crr` y `cuminc`. Así como algunas funciones derivadas de éstas.

4.2.1. Variable HP5

Para conocer la influencia que tiene o no la variable `HP5` (hipoxia), en la recaída local con o sin recaída distante de un paciente con cáncer de cuello uterino con el paquete `cmprsk`. Se utiliza la función `crr`, para llevar a cabo el análisis de datos con esta función, se tienen que construir algunas variables con los datos que contiene la base `ccup.csv`.

En la sección 3.1 se describieron las características que la función `crr` tiene, la sintaxis que se debe seguir y qué información aporta. Para trabajar con esta función, es necesario construir algunas variables. La primera es un vector que muestre los tiempos de falla/censura (en forma general se denómina como `ftime`). En este caso se utiliza la variable `dftime` contenida en la base de datos como vector de tiempos de falla/censura, y representa el tiempo del primer tipo de evento, ya sea: recaída local, recaída distante o ambas.

El segundo término que se debe construir es una variable que contenga cada tipo de falla y a la vez, pero por separado los casos censurados. Esta variable en la sintaxis general se expresa como `fstatus`, en este caso es nombrada como, `cen`. Dicho argumento debe contener un número que defina a la variable de interés y otro al riesgo competitivo. En este caso, el número 1 representa al evento de interés (recaída local con o sin recaída distante) y el número 2 describe al riesgo competitivo (recaída distante en ausencia de recaída local).

```
cen=(ccup$pelrec==‘‘Y’’)+2*(ccup$disrec==‘‘Y’’&
ccup$pelrec!=‘‘Y’’)
```

```
cen
  [1] 0 0 1 1 1 0 0 1 0 0 0 1 1 1 2 2 1 0 1 0 0 2 1 0
      0 0 1 0 1 2 1 0 0 1 0 2 1
 [38] 0 0 0 0 1 2 1 0 1 1 2 0 0 0 1 0 0 0 0 1 0 2 0 0
      2 0 2 2 0 2 1 0 1 2 0 1 2
 [75] 0 0 1 0 1 0 1 2 1 0 0 1 0 2 0 0 2 0 0 1 1 2 0 1
      0 0 1 0 0 0 0 0 0 0 0
```

La variable `cen` contiene 33 eventos de interés y 17 observaciones de riesgo competitivo, `cen=1= evento de interés`, `cen=2= riesgo competitivo` y `cen=0= censura`.

El tercer elemento que se debe construir es una matriz que contenga las covariables relacionadas con la variable de interés.

```
matricov=cbind(ccup$HP5,ccup$tamtum,(ccup$pelvi!=‘‘Y’’)+0)
```

La variable `matricov` es una matriz y su última columna un indicador con valor igual a 1, cuando el ganglio pélvico es negativo/equívoco y 0 cuando es positivo.

Con estos tres elementos es suficientes para aplicar la función `crr`. Al colocar `matricov[,1]`, definimos que el tamaño del tumor y el estado ganglionar no están controlados. El análisis se debe realizar por igual a los dos tipos de evento: interés y riesgo competitivo.

El primer evento analizado es el de interés (sin ajustar).

```
matricov=cbind(ccup$HP5,ccup$tamtum,(ccup$pelvicln!=‘‘Y’’)+0)
```

Se asigna un nombre a la variable que va a contener la información de `crr`, para este caso se nombra como `fit`:

```
fit=crr(ccup$dftime,cen,matricov[,1])
```

```
fit
convergence:TRUE
coefficients:
matricov[,1]1
0.007715
standard errors:
[1] 0.006191
two-sided p-values:
matricov[,1]1
0.21
```

Al aplicar la función `crr` se obtiene la siguiente información: convergencia, coeficiente, error estándar y p-value.

La hipótesis inicial de este análisis es que la hipoxia (`HP5`) influye en la recaída local de un paciente que ha recibido el tratamiento de quimioterapia y radioterapia. La hipótesis alternativa sería que la hipoxia no influye en la recaída local de dichos pacientes.

Dada la información, se observa que la convergencia se alcanza, el coeficiente es positivo, lo que significa que los valores más altos representan un mayor riesgo de recaída local. El `p-value=0.21`, esto quiere decir, que la variable `HP5`, no es significativa (considerando $\alpha = 0.05$).

En el siguiente escenario se controla el tamaño del tumor y el estado ganglionar. Para ello, sólo hay que retirar los corchetes `[,1]` de la variable `matricov`.

```
fit=crr(ccup$dftime,cen,matricov)
```

```
fit
convergence:TRUE
coefficients:
      HP5          tamtum          pelvi
  0.0001058    0.2883000    -0.6941000
standard errors:
[1] 0.006253 0.088440 0.415400
two-sided p-values:
   HP5     tamtum      pelvi
 0.9900    0.0011    0.0950
```

En este caso la función `crr` obtiene tres resultados representados en las columnas que corresponden al mismo orden de la matriz `matricov`: variable analizada (HP5), tamaño tumorial `tamtum` y estado ganglionar `pelvi`.

La información obtenida es la misma que en el primer caso: convergencia, coeficiente, error estándar y p-value. La influencia de las covariables es determinada por el p-value. El resultado de la variable hipoxia (`HP5`), nue-

vamente es no significativa, el tamaño del tumor (`tamtum`) es una covariable altamente significativa y el estado ganglionar (`pelvi`) es una covariable marginalmente signifitiva. Esto quiere decir, que las covariables: `pelvi` y `tamtum`, influye sobre `HP5`.

Cuando el modelo es ajustado por las covariables: `pelvi` y `tamtum`, se observa un cambio en el valor de p-value de la variable estudiada, en este caso es, `HP5`. En el primer escenario el `p-value=.21`, en el segundo es `p-value=.9900`, por lo general el valor incrementa en el caso donde se ajusta el modelo. Dicha variación ayuda a observar la influencia que ejercen las covariables sobre la variable `HP5`.

Con la información obtenida hasta ahora, aún no se puede concluir algo sobre la influencia de la variable `HP5` en la recaída local de un paciente. El mismo análisis que se realizó para el evento de interés se debe hacer para el evento riesgo competitivo.

En el riesgo competitivo, de igual forma, se analizan los dos casos: modelo sin ajustar (no se controlan las covariable tamaño tumorial y estado ganglionar) y modelo ajustado (se controlan las covariables tamaño tumorial y estado ganglionar). Dentro de la sintaxis de `crr` se debe aclarar que ahora se está trabajando con el riesgo competitivo, agregando la instrucción `failcode=2` dentro de `crr`.

Modelo no ajustado del riesgo competitivo.

```
fit=crr(ccup$dftime,cen,matricov[,1],failcode=2)
```

```
fit
convergence:TRUE
coefficients:
matricov[,1]1
0.01314
standard errors:
[1] 0.009166
two-sided p-values:
matricov[,1]1
   0.15
```

En este caso se puede obervar que existe la convergencia, el coeficiente es positivo. El `p-value=0.15`, muestra que la variable `HP5` no es significativa.

Ahora se analiza el riesgo competitivo en el modelo ajustado.

```
fit=crr(ccup$dftime,cen,matricov,failcode=2)
```

```
fit
convergence:TRUE
coefficients:
     HP5    tamtum      pelvi
 0.01587 -0.11080   0.32000
standard errors:
[1] 0.01011 0.14560 0.72920
two-sided p-values:
  HP5    tamtum      pelvi
 0.12     0.45       0.66
```

La convergencia existe en este caso también, el coeficiente es positivo. El `p-value=0.12`, implica que la variable `HP5` es no significativa. El tamaño del tumor, al igual que el estado ganglionar, son covariables no significativas, porque el `p-value` es mayor a 0.10 . En conclusión, la variable `HP5` (niveles bajos de oxígeno) no influye en la recaída local de un paciente con cáncer de cuello uterino primario. Esto quiere decir que los pacientes diagnósticados con cáncer de cuello uterino que recibieron tratamiento y presenten niveles bajos de `HP5`, tienen menor probabilidad de una recaída local.

4.2.2. Variable IFP

La segunda variable analizada es la presión del fluido intersticial `IFP`. La hipótesis con la que parte el análisis de esta variable es que la presión del fluido intersticial `IFP`, influye en la recaída local de un paciente con cáncer de cuello uterino primario. De igual forma se utilizan las herramientas del paquete `cmprsk` para analizar dicha variable.

Se utilizó la variable ya construida que contiene al evento de interés y el riesgo competitivo del caso `HP5`.

```
cen=(ccup$pelrec==‘‘Y’’)+2*(ccup$disrec==‘‘Y’’&
ccup$pelrec!=‘‘Y’’)
```

```
cen
  [1] 0 0 1 1 1 0 0 1 0 0 0 1 1 1 2 2 1 0 1 0 0 2 1
      0 0 0 1 0 1 2 1 0 0 1 0 2 1
 [38] 0 0 0 0 1 2 1 0 1 1 2 0 0 0 1 0 0 0 0 1 0 2 0
      0 2 0 2 2 0 2 1 0 1 2 0 1 2
 [75] 0 0 1 0 1 0 1 2 1 0 0 1 0 2 0 0 2 0 0 1 1 2 0
      1 0 0 1 0 0 0 0 0 0 0 0
```

La variable `cen` muestra los eventos censurados y tipos de falla.

Se construyó la matriz de covariables tomando en cuenta las mismas variables que en el caso de la variable `HP5`.

```
matricov=cbind(ccup$IFP,ccup$tamtum,(ccup$pelvicln!=‘‘Y’’)+0)
```

Con los elementos necesarios se aplica la función `crr`, al primer escenario de la variable de interés, modelo sin ajustar.

```
fit=crr(ccup$dftime,cens,matricov[,1])
```

```
convergence:TRUE
coefficients:
matricov[,1]1
0.03441
standard errors:
[1] 0.01699
two-sided p-values:
matricov[,1]1
   0.043
```

Se obtiene la convergencia, el coeficiente, el error estándar y el p-value. El resultado indica que la convergencia se alcanzó y el `p-value=0.043` de `IFP` es significativo. El coeficiente es positivo e implica que los valores más altos de `IFP` representan un mayor riesgo para que un paciente sufra una recaída local.

Se utiliza la función `crr` para observar el modelo ajustado.

```
fit=crr(ccup$dftime,cen,matricov)
```

```
fit
convergence:TRUE
coefficients:
     IFP      tamtum       pelvi
 0.03247    0.26110    -0.80790
standard errors:
[1] 0.01705 0.09664 0.42920
two-sided p-values:
   IFP    tamtum     pelvi
0.0570    0.0069    0.0600
```

El valor de los coeficientes, error estándar y los p-value tienen el mismo orden que la matriz `matricov`. El p-value para `IFP=0.057`, quiere decir que la variable `IFP` es marginalmente significativa. El `p-value` de la covariable `tamtum`, es altamente significativo. La covariable `pelvi` es marginalmente significativa. Esto implica, que las variables: `tamtum` y `pelvi`, influyen en la variable `IFP`.

El resultado obtenido muestra que al ajustar la variable `IFP` por las covariables: tamaño del tumor y estatus ganglionar. El `p-value` de `IFP` incrementa, convirtiéndose de significativo a marginalmente significativo, porque, $0.01 \leq$ `p-value` < 0.05 (significativo) y $0.05 \leq$ `p-value` < 0.10 (marginalmente significativo).

De igual forma se ejecuta el mismo modelo para el primer riesgo competitivo (recaída distante en ausencia de recaída local), específicando en la sintaxis que ahora se está trabajando el caso 2, `failcode=2`, modelo no ajustado.

```
fit=crr(ccup$dftime,cen,matricov[,1],failcode=2)
```

```
fit
convergence:TRUE
coefficients:
matricov[,1]1
0.04377
standard errors:
[1] 0.01955
two-sided p-values:
matricov[,1]1
  0.025
```

Se obtiene la misma información que en el modelo no ajustado, pero para el primer caso de riesgo competitivo (recaída local en ausencia de recaída distante). Se puede observar que la variable `IFP` es significativa con un `p-value=0.025`, esto quiere decir, que la variable `IFP` influye en la recaída local.

Posteriormente, se construye el modelo para el caso ajustado (control en el tamaño del tumor y el estado ganglionar).

```
fit=crr(ccup$dftime,cen,matricov,failcode=2)
```

```
fit
convergence:TRUE
coefficients:
```

```
     IFP       tamtum        pelve
 0.05167    -0.15470    -0.12490
standard errors:
[1] 0.02074 0.16710 0.65540
two-sided p-values:
   IFP  tamtum  pelve
0.013  0.350   0.850
```

La variable `IFP` en este caso sigue siendo significativa con un `p-value=0.013`. Las covariables tamaño del tumor y el estado ganglionar, no son significativas, esto quiere decir, que no influyen en la variable `IFP` en el modelo ajustado.

El análisis aún no ha concluído, debido a que hay que estudiar la segunda alternativa del riesgo competitivo (recaída local con recaída distante). Para realizarlo, es necesario modificar la variable que contiene al evento de interés y aplicar de nuevo la función `crr`, al riesgo competitivo.

El análisis tiene como objetivo estudiar la influencia de las variables: `HP5` e `IFP` en la recaída local. Los resultados de la primera variable (`HP5`) mostraron que dicha variable no es significativa, esto quiere decir, que no influye en la recaída local de los pacientes después de recibir un tratamiento de quimioterapia y radioterapia. Por lo tanto, el evento de interés (recaída local con o sin recaída distante) no se observó. El análisis de la variable `HP5`, concluye porque el evento de interés no sucedió.

En el caso de la variable `IFP` el evento de interés se observó y el análisis aún no término, porque se tiene que establecer si esta variable influye en la recaída local del evento alterno (riesgo competitivo). Para poder analizar el segundo evento, la sintaxis de la función tiene que ser modificada. En la hipótesis original (primer caso), el evento de interés omitía la posibilidad de recaída local, lo cual no puede ser afirmado, debido a que no se encontró evidencia suficiente para sustentar dicha afirmación, por la no significancia de las covariables. Por ello, es necesario que el evento alterno (riesgo competitivo) cambio a recaída distante con recaída local.

Es importante señalar que aunque se trate de un análisis de riesgos competitivos la lógica del estudio debe respetar la hipótesis original, la cual afirma la influencia de las variables `IFP` y `HP5`, en la recaída local de los pacientes que recibieron un tratamientos contra el cáncer uterino. Por ejemplo, hubiera sido un error analizar como riesgo competitivo la recaída distante, porque la hipótesis original habla sobre recaída local, y la recaída distante no es un riesgo competitivo, ya que el significado de ésta es el mismo en ambos escenarios: recaída local y distante (reincidencia de la enfermedad en un lugar lejano al cuello uterino). El interés de estudiar la recaída local surge porque puede significar dos cosas totalmente distintas, la primera es que el tumor nunca desapareció, y la segunda que el cáncer reincidió en un lugar cercano al de origen, la pelvis.

Para realizar el análisis de riesgo competitivo se construye una variable que contenga al evento alterno (riesgo competitivo), recaída distante con recaída local.

```
cens=(ccup$disrec==‘‘Y’’)+2*(ccup$disrec==‘‘N’’&
ccup$pelrec==‘‘Y’’)
```

```
cens
  [1] 0 0 2 1 2 0 0 1 0 0 0 1 2 1 1 1 2 0 1 0 0 1 2 0
      0 0 1 0 2 1 1 0 0 2 0 1 1
 [38] 0 0 0 0 2 1 1 0 2 2 1 0 0 0 1 0 0 0 0 1 0 1 0 0
      1 0 1 1 0 1 1 0 2 1 0 2 1
 [75] 0 0 2 0 1 0 1 1 1 0 0 2 0 1 0 0 1 0 0 2 2 1 0 2
      0 0 2 0 0 0 0 0 0 0 0
```

En este caso se analizan los casos igual a 2.

Se utiliza la misma matriz de covariables `matricov`, porque se deben utilizar los mismo elementos que se emplearon en el análisis del evento de interés.

```
matricov=cbind(ccup$IFP,ccup$tamtum,(ccup$pelvi!=‘‘Y’’)+0)
```

Se aplica las función `crr`.

```
fit=crr(ccup$dftime,cens,matricov)
```

```
fit
convergence:TRUE
coefficients:
    IFP      tamtum       pelve
 0.04029    0.14780    -0.29130
standard errors:
[1] 0.01847 0.10720 0.47960
two-sided p-values:
 IFP   tamtum    pelve
0.029   0.170   0.540
```

El `p-value=0.029`, indica que la variable `IFP` es significativa, pero las covariables `tamtum` y `pelve`, no son significativas. Con dicha información no se puede obtener una conclusión sobre si la variable `IFP` influye en la recaída local con recaída distante. Por ello es necesario construir grupos que contenga la variable de interés (IFP) junto con las covariables y comparar distintos escenarios.

Con la función `predict.crr` se puede calcular la estimación de la causa-específica del riesgo acumulado. El primer parámetro de la función es un objeto, y es la salida de la función de `crr`. Los parámetros dos y tres (cov1 y cov2), son matrices de covariables y deben tener la misma estructura que sus homólogos utilizados en la función `crr`. Si las covariables elegidas no dependen del tiempo, basta con construir únicamente una covariable, `cov1`.

La hipótesis dice que niveles elevados de `IFP` influyen en la recaída distante (RC). Los escenarios utilizados por recomendaciones médicas, son: nivel elevado de `IFP` (30) y nivel bajo de `IFP` (5). Estos casos toman en cuenta la influencia del tamaño tumorial y el estado ganglionar. El tamaño del tumor (5 cm), es el mismo para los cuatro escenarios, esto es necesario para que puedan ser comparables. El estado ganglionar tiene dos posibles respuesta: *negativo/equívoco* = 1 y *positivo* = 0.

1. IFP = 5, Tamaño del tumor= 5, Ganglio Negativo/Equívoco= 1.
2. IFP = 5, Tamaño del tumor= 5, Ganglio Positivo = 0.

3. IFP = 30, Tamaño del tumor= 5, Ganglio Negativo/Equívoco= 1.

4. IFP = 30, Tamaño del tumor= 5, Ganglio Positivo= 0.

Se construye la matriz `cov1`, con los escenarios de `IFP`, los cuales se determinan por el tamaño del tumor y el estadio ganglionar.

```
pfit=predict.crr(fit,px)
```

```
pfit
      IFP tamtum pelvi
[1,]   5      5     1
[2,]   5      5     0
[3,]  30      5     1
[4,]  30      5     0
```

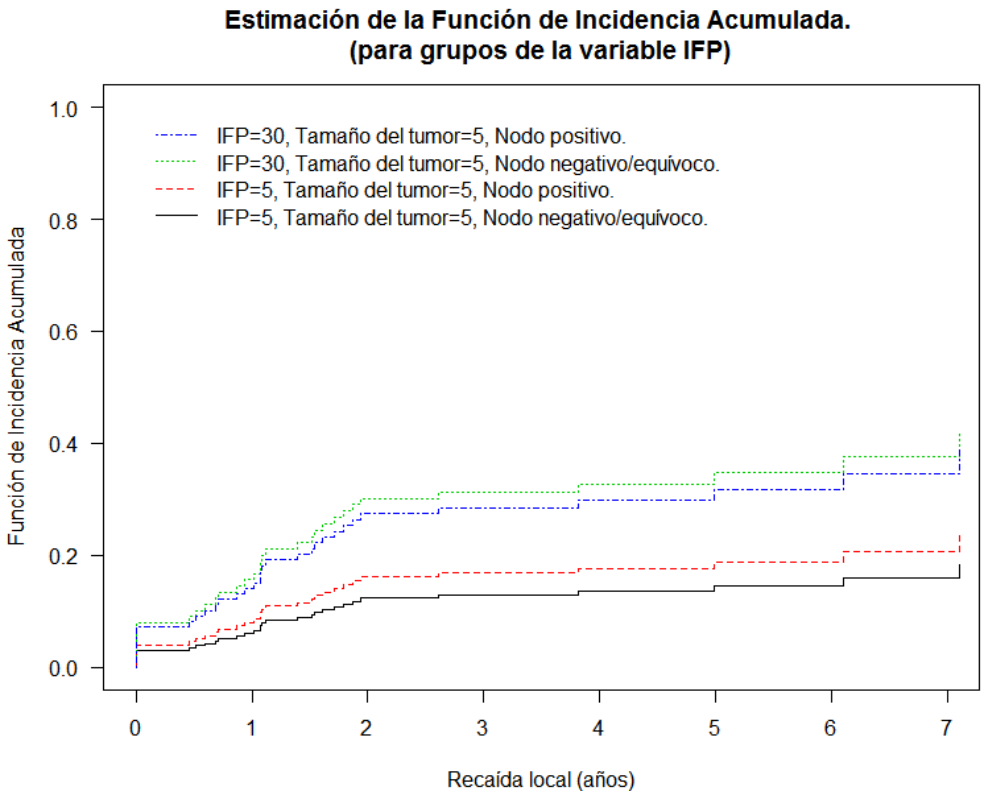
El número de filas corresponde a los escenarios y el de columnas al número de covariables.

Cada fila es un conjunto de valores de las covariables para las que se desea calcular la función de incidencia acumulada. El resultado es una matriz con tantas filas como el número de tiempos de falla. El número de columnas es igual al número de filas de `cov1` más uno para el punto de tiempo.

Para observar el comportamiento de la función `predict.crr`, se utiliza la función `plot.predict.crrr`, la cual construye un histograma y muestra el comportamiento de los cuatro escenarios que contiene la variable `pfit`.

```
par(las=1,mfrow=c(1,1))
plot.predict.crr(pfit,lty=c(1:4),ylim=c(0,1),color=1:4,
xlab="Recaida local (anos)", ylab=``Funcion de
Incidencia Acumulada", main=``Estimacion de la Funcion
de Incidencia Acumulada.(para grupos de la variable
IFP)", col.main=``black") legend(0,1,lty=c(1:4),
bty=``n", legend=c(``IFP=5, Tamano del tumor=5,
```

```
Nodo negativo/equivoco.", ``IFP=5, Tamano del tumor=5,
Nodo positivo.", ``IFP=30, Tamano del tumor=5,
Nodo negativo/equivoco.",``IFP=30, Tamano del tumor=5,
Nodo positivo."))
```



La gráfica anterior muestra la comparación de los escenarios de riesgos competitivos. El primer caso muestra un nivel de `IFP` igual a 5, influido por las covariables: tamaño tumoral y ganglio. En el segundo caso se toma en cuenta el nivel de `IFP` igual a 30. Cabe notar que la covariable tamaño tumorial es fija, la covariable que tiene dos posibles resultados, es el estadio ganglionar: positivo o negativo.

Cuando el nivel de `IFP`, es bajo, y se tiene un ganlio/nodo positivo, la probabilidad de sufrir una recaída local con recaída distante es mayor (línea roja), en comparación con aquellos pacientes que tienen el mismo nivel de

`IFP` pero con ganglio/nodo negativo (línea negra).

Para el caso cuando los niveles de `IFP` son altos, en presencia de ganglio/-nodo negativo la probabilidad de presentar una recaída local con recaída distante (línea verde) es más alta que aquellos pacientes con elevado `IFP` con ganglio/nodo positivo (línea azul).

En la gráfica 4.2 también se puede observar que niveles elevados de líquido intersticial (línea verde y azul), influyen en la recaída local con recaída distante en pacientes con cáncer de cuello uterino primario que recibieron tratamiento. Esto quiere decir, que un paciente con elevado `IFP` tiene mayor probabilidad de tener reincidencia de cáncer en la pélvis, con o sin ganglio/-nodo positivo o negativo.

La función de incidencia acumulada se puede obtener a partir de dos funciones del paquete `cmprsk`: `predic.crr` y `cuminc`. Para observar la construcción y diferencia entre ambas se utilizaron los datos del caso riesgo competitivo de la variable `IFP`.

La construcción a partir de la función `predict.crr`, parte de la función `crr`. En este caso, se calcula $\{1 - e^{-B(t)}\}$, donde B(t) es la estimación del riesgo de la causa-específica acumulada que se obtiene para los valores de las covariables específicas a partir de la estimación de tipo Breslow[3] del riesgo base y los coeficientes de regresión estimados.

El siguiente modelo analiza la incidencia acumulada de pacientes con ganglio negativo/equívoco o ganglio positivo, en presentar una recaída local con recaída distante. Se estudia esta covariable, porque varía para determinar su influencia en la variable `IFP`, a diferencia de la covariable `tamtum`, que en análisis anterior de bajos y altos niveles de `IFP`, el tamaño tumoral se tomó como una covariable fija.

Se utiliza la función `crr`, la cual requiere de tres parámetros: vector de tiempo `dftime`, variable `cens` y una matriz de covariables `matripelve`.

[3] 1970. Kruskal-Wallis.

En este caso la variable `matripelve`, difiere de los anteriores casos, porque únicamente se toma en cuenta una covariable, estado ganglionar (`pelvi`), la cual tiene dos posibles escenarios: ganglio negativo/equívovo o ganglio positivo. Con este modelo se puede observar el comportamiento de la incidencia acumulada de la variable `IFP` en ambos escenarios: ganglio/nodo positivo y negativo.

La construcción del vector `matripelve`, que contiene la covariables analizada es:

```
matripelve=1-(ccup$pelvi!=``Y'')+0
```

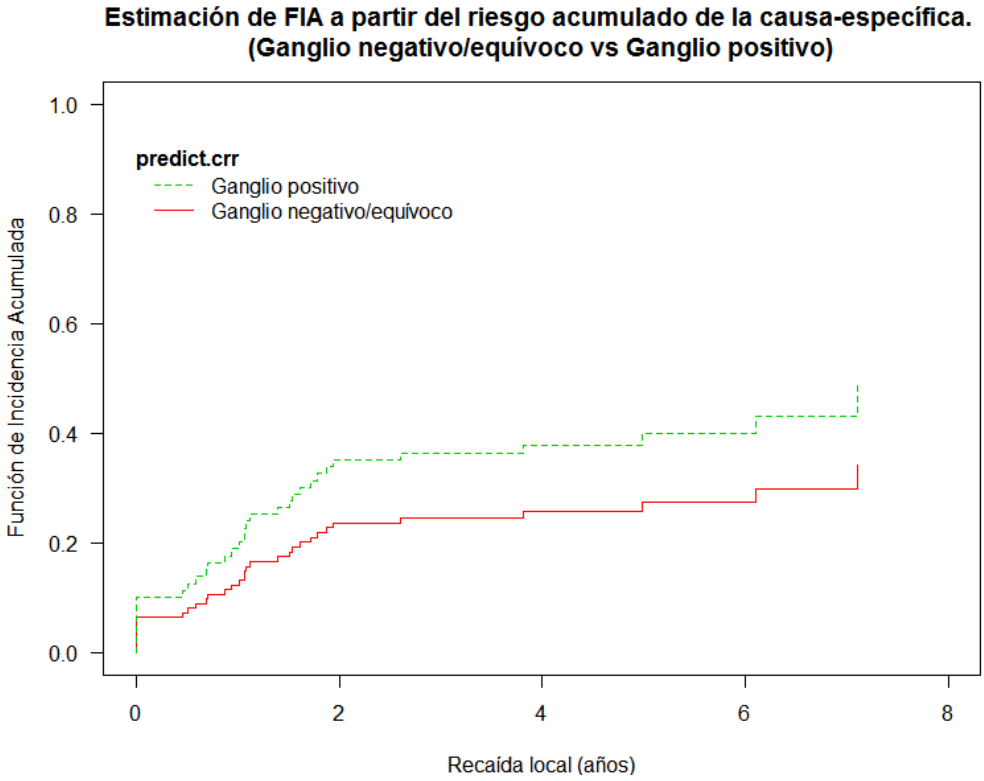
```
matripelve
  [1] 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0
      1 0 0 0 1 1 0 0 0 1 0 0 0
 [38] 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 1 1 1 0 0 0 0
 [75] 0 1 0 0 1 0 1 0 1 1 1 0 0 1 0 0 0 0 0 0 0 0 0 0 1
      0 0 0 1 0 0 0 0 1 0 0
```

El número 0, indica ganglio negativo/equívoco, y el número 1, ganglio positivo. El análisis de riesgos competitivos bajo el enfoque de la función de incidencia acumulada permite estudiar a más de un escenario a la vez.

Se aplica la función `crr`, con la nueva matriz de covariable `matripelve`.

```
cens=(ccup$disrec==``Y'')+2*(ccup$disrec==``N''&
ccup$pelrec==``Y'')
matripelve=1-(ccup$pelvi!=``Y'')+0
fit=crr(ccup$dftime,cens,matripelve)
px=matrix(c(0,1),ncol=1)
pfit=predict.crr(fit,px)
par(las=1,mfrow=c(1,1))
plot.predict.crr
```

Con la función `plot.predict.crr`, se construye el gráfico que muestra el comportamiento de ambos escenarios (ganglio negativo/equívoco y ganglio positivo).



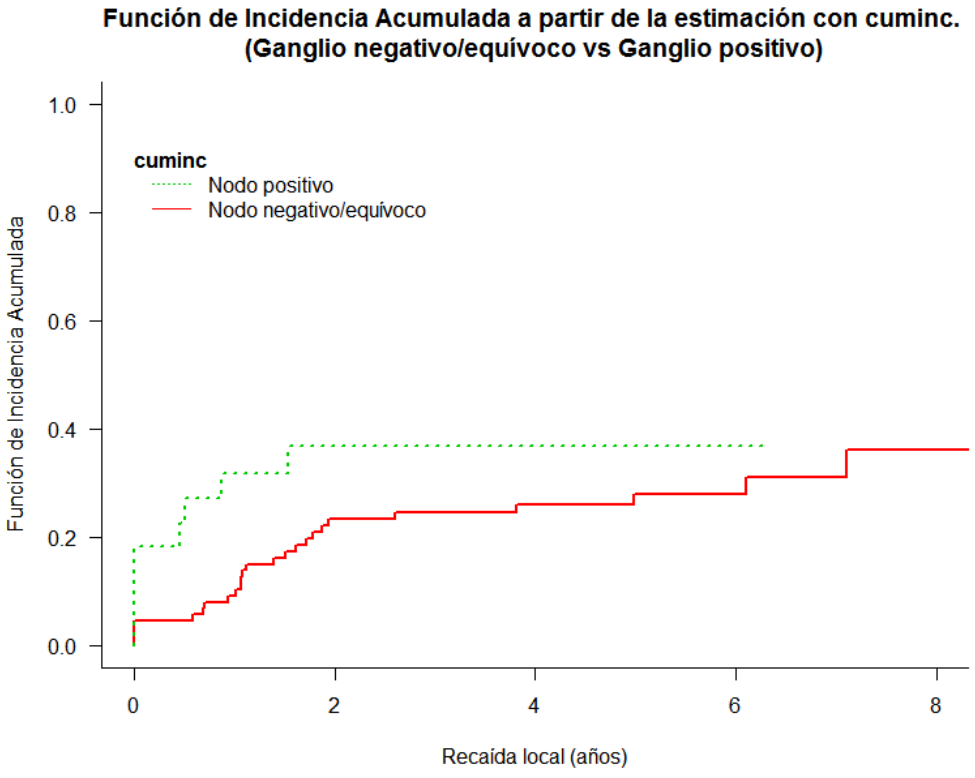
La gráfica anterior muestra que la incidencia acumulada de recaída local, construída con la acumulación del riesgo de causa-específica, es mayor cuando el paciente tiene un ganglio positivo. La diferencia entre ambos escenarios es considerable, ya que poco después de algunos meses se puede observar que la incidencia acumulada aumenta para el caso de ganglio positivo.

Ahora, se estudiarán los mismos datos pero desde el enfoque de la función `cuminc`. La cual estima las funciones de incidencia acumulada de los datos que son riesgo competitivo y prueba la igualdad entre los grupos. La construcción de la función `cuminc`, requiere de los mismos parámetros utilizados en la función `crr`: tiempo de falla, variable que contiene los tipos de falla/-censuras y la matriz de covariables.

Se construye una variable que contenga las 3 salidas de `cuminc`, en este caso se llamará `fitc`:

```
fitc=cuminc(ccup$dftime,cens,matripelve)
forplot=list(list(fitc$``0 1''$time,fitc$``0 1''$est),
list(fitc$``1 1''$time,fitc$``1 1''$est))
```

Para construir la gráfica de la función `cuminc`, se utiliza `plot.cuminc`, la cual se construye de igual forma que `plot.predict.crr`.



La gráfica anterior muestra que es mayor la incidencia acumulada en la recaída local, si el paciente tiene un ganglio positivo, pero la gráfica también muestra que a partir de cierto tiempo la incidencia de ambos casos se iguala.

En la gráfica obtenida con la función `plot.predict.crr`, la diferencia entre ambos escenarios (ganglio negativo/equívoco y ganglio positivo) es clara, el gráfico muestra que la incidencia acumulada con ganglio positivo es mayor que la incidencia con ganglio negativo/equívoco.

La mayor incidencia acumulada de recaída local tiene el mismo comportamiento en ambos escenarios, aunque hay una diferencia de comportamiento a largo plazo. La probabilidad de observar una recaída local en presencia de recaída distante es más alta en pacientes que dieron positivo en ganglio, que aquellos con ganglio negativo/equívoco. Aunque en ambos casos: `predict.crr` y `cuminc`, se obtuvo el mismo resultado.

La función `predict.crr`, muestra variaciones en el nivel de incidencia acumulada, mientras que la función `cuminc`, tiene un comportamiento constante que se iguala al final de los años. Utilizar una u otra herramienta depende de quién esté elavorando en análisis. Lo más recomendable es que se utilicen ambas, para corrrobar que el comportamiento de la variable sea el mismo con ambas funciones.

4.3. Simulación

En esta sección se construyen un par de simulaciones para mostrar las dos principales funciones del paquete `cmprsk`, que obtienen la función de incidencia acumulada: `crr` y `cuminc`. Determinan a través de una comparación qué variable tiene mayor incidencia. Así como las covariables que establecen esa mayor incidencia de la variable.

Los primero que se debe determinar es la semilla, la cual permite que los datos simulados sean los mismos bajo cualquier función que se utilice. Después se crea la variable tiempo, en este caso, se indicó que se simularán 200 distribuidos exponencialmente[4]. Después se construye la variable status, se utiliza la función `sample`, la cual toma una muestra del tamaño de la simulación (en este caso 200), ya sea con o sin sustitución. Posteriormente se construye la matriz de covariables, indicando el número de covariables que se van a utilizar.

Con las tres variables construidas: tiempo, status y matriz de covariables, ya se puede aplicar la función `crr` a los datos.

```
set.seed(15)
f.time<-rexp(200)
fstatus<-sample(0:2,200,replace=TRUE)
cov<-matrix(runif(600), nrow=200)
dimnames(cov)[[2]]<-c('ftime','fstatus','cov')
print(z<-crr(ftime,fstatus,cov))
summary(z)
```

```
convergence:  TRUE
coefficients:
  ftime   fstatus   cov
-0.2738  -0.0141  -0.3781
standard errors:
[1] 0.4163 0.4370 0.4434
two-sided p-values:
  x1    x2    x3
0.51  0.97  0.39
```

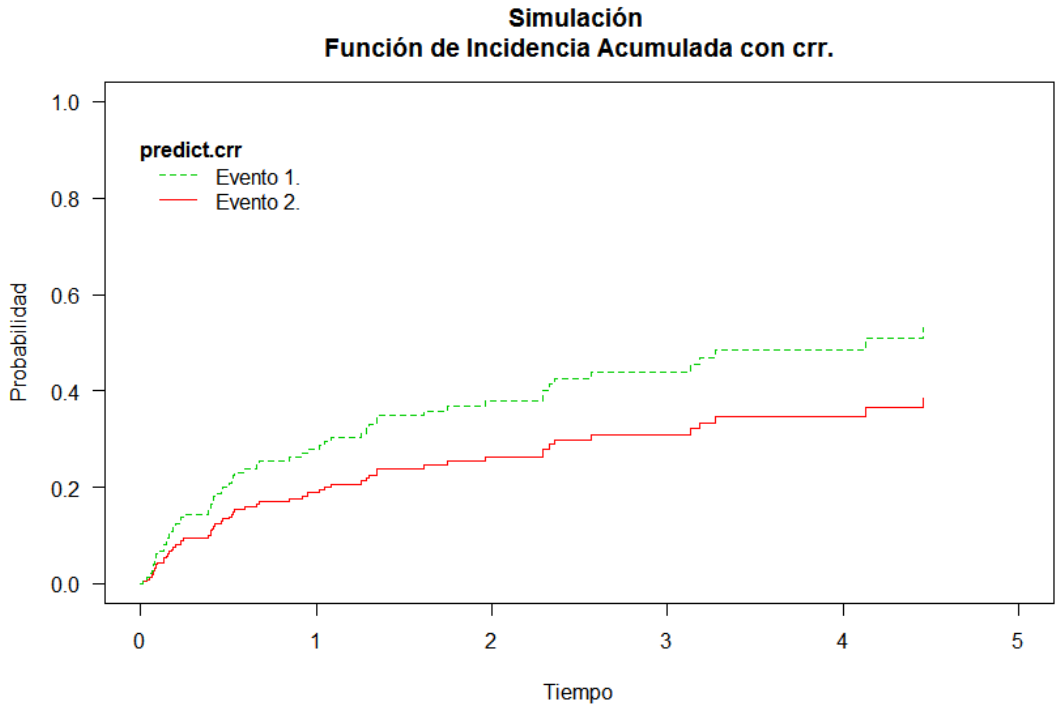
[4]La elección del número de simulaciones es arbitraria.

La información obtenida son los `p-value` de las variables construidas: tiempo, status y matriz de covariables, lo cual permite determinar el nivel de significancia.

Para comparar dos eventos es necesario incluir a ambos mediante la función `rbind`, la cual hace coincidir las filas y columnas de una matriz, para que puedan ser comparables.

Finalmente, se utiliza la función `plot`, para obtener la gráfica.

```
z.p<-predict.crr(z,rbind(c(.1,.5,.8),c(.1,.5,.2)))
plot(z.p,lty=c(1:2),color=2:3, ylim=c(0,1), xlim=c(0,5),
xlab="Tiempo", ylab="Probabilidad",main="Simulacion de
la Funcion de Incidencia Acumulada con crr.")
text(0,.9,adj=0,"predict.crr", font=2)
legend(0,.9,lty=c(2:1),col=c(3,2), bty="n",
legend=c("Evento 1.","Evento 2."))
```

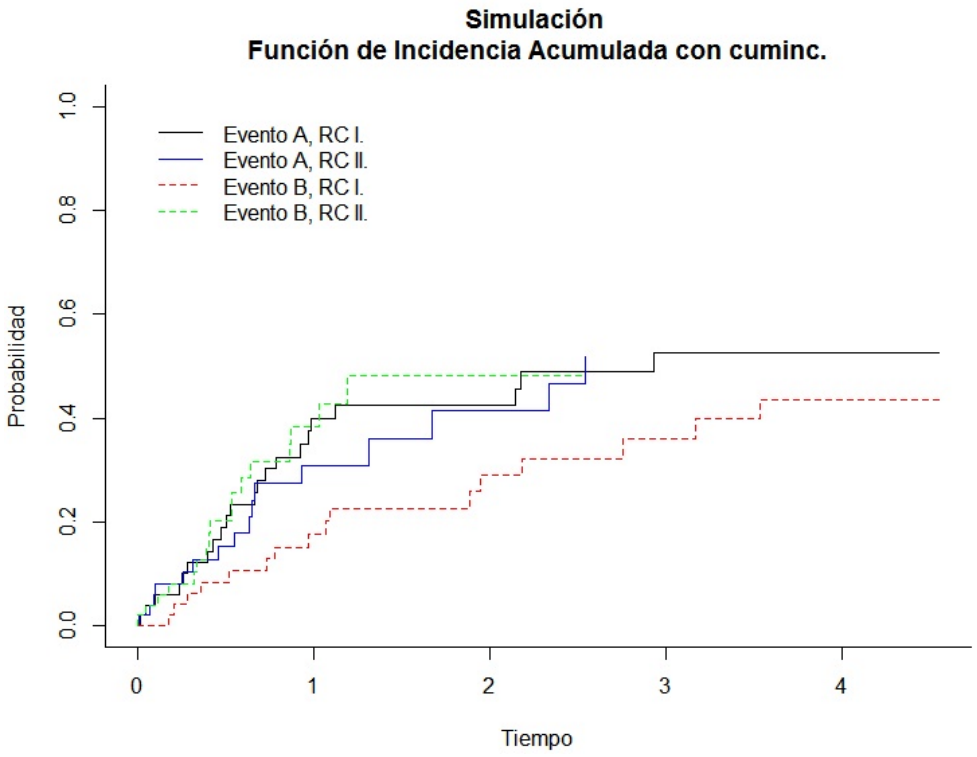


La gráfica anterior muestra la incidencia acumulada de cada evento. El evento número 1 tiene mayor probabilidad de suceder (línea verde) en comparación con el evento número dos (línea roja). También se puede notar que la diferencia de probabilidad entre ambos eventos se presenta desde el inicio del tiempo.

Ahora se construye una simulación para utilizar la función `cuminc`. Con ella, se pueden analizar los eventos agrupados bajo características específicas. Requiere de tres vectores: tiempo, status y vector que indique un subcobjunto.

El primer paso de la simulación es plantar una semilla, se necesita crear un conjunto de datos, en este caso son 100, que se distribuyen exponencialmente, también es necesario construir un vector tiempo. Se deben agrupar en variables los casos que se van a comparar.

```
set.seed(15)
ss <- rexp(100)
gg <- factor(sample(1:2,100,replace=TRUE),1:2,
c('Evento A','Evento B'))
cc <- sample(0:2,100,replace=TRUE)
strt <- sample(1:2,100,replace=TRUE)
print(xx <- cuminc(ss,cc,gg,strt))
plot(xx, lty = c(1, 1, 2, 2),
col=c("black","blue","red","green"),
curvlab=c("Evento A,RC I.", "Evento A, RC II.",
"Evento B,RC I.","Evento B, RC II."),
xlab="Tiempo",ylab="Probabilidad",
main="Simulacion Funcion de Incidencia
Acumulada con cuminc.")
```



En la gráfica anterior se puede observar la comparación de dos eventos en situación de riesgos competitivos. Los datos se agruparon para el riesgo competitivo I y II, para ambos eventos (A y B).

El evento A con riesgo competitivo I, tienen mayor probabilidad de ocurrir que el evento A bajo riesgo competitivo II. En el caso del evento B, bajo riesgo competitivo I, es menor la probabilidad de que suceda, en comparación con el riesgo competitivo II, que tiene mayor probabilidad de suceder.

Conclusiones

El presente trabajo es una introducción al análisis de riesgos competitivos bajo el enfoque de la función de incidencia acumulada. El objetivo principal de su elaboración fue exponer en que consiste dicho análisis. Así como mostrar las herramientas de la paquetería `cmprsk` del programa `r.project`, que se pueden utilizar para construir un análisis de riesgos competitivos más completo.

El análisis de riesgos competitivos pertenece al análisis de supervivencia, pero este último únicamente se enfoca en estudiar un sólo evento de interés. El análisis de riesgos competitivos, amplía el panorama de estudio del de supervivencia, ya que permite estudiar eventos alternos, a través de covariables, para obteneter más de un escenario, y con ello, construir conclusiones pertinentes que ayuden a mejorar tratamientos, procesos, ciclos, etc.

Actualmente, la recolección de información que construye las bases de datos, utilizadas para diversos análisis médicos o de cualquier tipo, se elaboran con base en amplios seguimientos de diversas variables de valoración, con el fin de que la multiplicidad de eventos que puede presentar un paciente bajo estudio o un proceso, puedan ser examinadas con riesgos competitivos y estudiar todos aquellos casos que dentro del análisis de supervivencia se consideran censura.

Con el análisis realizado en este trabajo a la base de datos `ccup.csv`, que contiene información sobre pacientes diagnósticadas con cáncer de cuello uterino en su fase primaria, a través del análisis de riesgos competitivos, con la función de incidencia acumulada, se obtuvieron conclusiones sobre la influencia de las variables: `HP5` e `IFP`, en la recaída local de los pacientes con cáncer de cuello primario que recibieron tratamiento de quimioterapia

y radioterapia.

Los resultados obtenidos con las funciones del paquete `cmprsk`, indican que niveles bajos de hipoxia (disminución en la cantidad de oxígeno suministrado por la sangre a los orgános), no influye en la recaída local de un paciente que padeció cáncer de cuello uterino, fase primaria. Se hicieron dos escenarios: modelo ajustado y modelo no ajustado. En el primer caso únicamente se analizó la variable `HP5`. En el segundo modelo se agregaron las covariables: tamaño del tumor y estadio ganglionar, se encontró que dichas covariables no influyen en la variable `HP5`, y por lo tanto, no modifican la probabilidad de recaída local de un paciente. Esto quiere decir, que todas aquellas pacientes que fueron diagnosticadas con la enfermedad de cáncer uterino en la fase primaria, que recibieron tratamiento y presentan niveles bajos de hipoxia (`HP5`), no tienen mayor probabilidad de sufrir una recaída local.

La segunda variable que se analizó fue `IFP` (líquido intersticial, el cual sirve para rellenar la parte vacía entre las células y los capílares sanguíneos) y se determinó la influencia que esta variable tiene en la recaída local de una paciente con cáncer uterino. Los resultados obtenidos indican que niveles elevados de `IFP`, sí influyen en la recaída local de una paciente con cáncer uterino (fase primaria) que recibió tratamiento. Al realizar el análisis de la variable `IFP`, se encontró a través de los modelos: ajustado y no, es significativa y marginal. Esto quiere decir, que sí influye en la recaída local de pacientes diagnósticadas con cáncer de cuello que recibieron quimioterapia y radioterapia.

Los primeros escenarios analizados de la variable `IFP`, mostraron que sí influye en la recaída local en ausencia de recaída distante. El análisis de riesgos competitivos estudia todos los posibles escenarios. En este caso, el segundo es la recaída local con recaída distante y también se utilizaron covariables para determinar la influencia que éstas tienen en la variable. Las covariables utilizadas fueron las mismas que en análisis anteriores: tamaño del tumor y estado ganglionar. Cabe señalar que en éste y en todo análisis de supervivencia es muy importante no perder la lógica del estudio.

Para determinar la influencia de la variable `IFP` en la recaída local con distante, fue necesario establecer parámetros para poder comparar los escenarios.

En el primer caso se analizó la influencia de la covariable estado ganglionar (`pelvi`), en sus dos estados: equívoco y positivo, junto con la covariable tamaño tumoral (variable fija). Por sugerencia médica se tomó como nivel elevado `IFP=30` y bajo `IFP=5`. En este caso se construyó una función que contuvo cuatro escenarios. Como resultado se obtuvo, que niveles elevados de líquido intersticial influyen en la recaída local con distante en mujeres con cáncer de cuello uterino (fase primaria) que recibieron tratamiento de quimioterapia y radioterapia. Esto quiere decir, que aquellos pacientes con elevado índice de `IFP`, tienen mayor probabilidad de padecer una reincidencia de cáncer en la pelvis.

Como la variable estado ganglionar tiene dos posibles resultados: negativo/equívoco y positivo. Se analizó la incidencia acumulada que cada uno de estos escenarios tiene. El análisis de la función de incidencia acumulada se hizo a través de dos funciones: `predict.crr` y `cuminc`.

Con las dos funciones se obtuvo el mismo resultado, la incidencia acumulada es mayor en el caso de ganglio positivo. Existe una diferencia entre utilizar la función `predict.crr` y `cuminc`. Con la gráfica de la función `predict.crr` se puede observar que a través del tiempo, la incidencia acumulada siempre es mayor cuando el ganglio es positivo. Cuando se utiliza la función `cuminc`, al inicio la función de incidencia acumulada es mayor de igual forma en el caso de ganglio positivo, pero al pasar el tiempo ambos escenarios se igualan. Decidir qué método es más conveniente utilizar depende de varios factores, como las características de los datos y los objetivos de cada investigador. Lo que se recomienda con este trabajo es construir ambas funciones, con el fin de comparar que con ambas se llega al mismo resultado.

Con los ejemplos desarrollados en la sección de simulación, de forma un poco más sencilla se mostró la construcción de las funciones `crr` y `cuminc`. Con la primera se obtienen una comparación entre dos eventos, la gráfica muestra cual tiene una mayor incidencia acumulada.

La función `cuminc`, permite analizar varios eventos a través de agrupaciones dentro de las propias variables, compara diversos escenarios a la vez. En la simulación (pág. 71), se puede observar la gráfica de dos eventos, los cuales están bajo dos riesgos competitivos. Este tipo de análisis pérmite estudiarlos a la par, lo cual es de máximo valor para la construcción de conclusiones.

El paquete `cmprsk` se compone de diversas funciones que modelan el comportamiento de datos que se encuentran bajo el análisis de riesgos competitivos. A través del análisis de covariables, las cuales ayudan a determinar si la variable estudiada influye o no en el evento de interés y en el riesgo competitivo. El objetivo de utilizar esta paqueteria es fomentar su uso y dejar abierta la expansión de este tema.

Apéndice A. Estimador de Kaplan Meier y riesgos proporcionales de Cox.

Kaplan-Meier

El estimador de Kaplan-Meier calcula la supervivencia cada vez que se presenta una falla, este método obtiene proporciones exactas de supervivencia porque utiliza tiempos precisos, también es conocido como *producto límite*.

La proporción acumulada que sobrevive se calcula para el tiempo de supervivencia individual de cada sujeto y no se agrupan los tiempos de supervivencia en intervalos. Por esta razón es especialmente útil para análisis que tienen pocos datos. Kaplan-Meier incorpora la idea del tiempo en que ocurren los eventos, su estructura es la siguiente:

$$\widehat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right) \quad \text{donde} \quad t_k \leq t \leq t_{k+1}$$

- n_j**.** Representa el número de sujetos en riesgo, es decir, que no han presentado la falla antes del tiempo $(t_{j-\delta})$ con $(\delta > 0)$ y δ lo suficientemente pequeña.
- $(t_{j-\delta})$ Es un instánte del tiempo de interés y es ahí donde se observa cuántos individuos se tienen hasta ese momento.
- d_j Representa el número de fallas al tiempo t_j.

El estimador de Kaplan-Meier se utiliza para describir el tiempo transcurrido hasta que se presenta el evento de interés. Tiene ventajas sobre otros métodos de análisis: es fácil de calcular, se pueden obtener gráficos que muestran de forma clara los resultados y su interpretación resulta fácil.

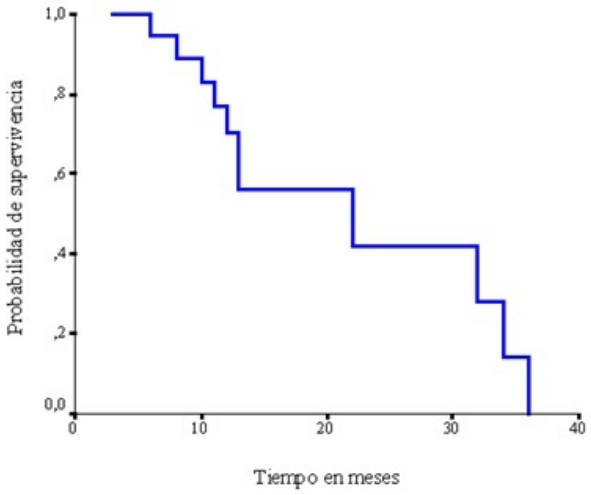


Figura 4.1: Estimador de Kaplan-Meier

La figura 4.1 muestra el comportamiento general del estimador de Kaplan-Meier, se puede observar la probabilidad de supervivencia en el tiempo establecido. En este caso la escala está definida en meses, la cual se puede establecer en diversas: años, días, horas, minutos, etc.

El método más frecuente para estimar la probabilidad de un evento es un enfoque no paramétrico, generalmente denominado método de Kaplan-Meier (KM) o método de límite de producto. El supuesto principal de la estimación de KM de la supervivencia es que las observaciones censuradas acabarían presentando el evento si el seguimiento fuera lo bastante largo.

Riesgos proporcionales de Cox

Existen diversos modelos de supervivencia que involucran covariables al incorporar la manera en que éstas afectan el tiempo de falla del sujeto bajo estudio. Resulta interesante poder modelar no sólo la relación entre la tasa de supervivencia y el tiempo, sino también la posible relación con diferentes variables registradas para cada sujeto, esto quiere decir, que se puede calcular la tasa de supervivencia como una función del tiempo y de las variables pronóstico (covariables).

El modelo que puede definir la relación ente la función de supervivencia y las covariables dentro del análisis de supervivencia es el de riesgos proporcionales de Cox. Dicho modelo es el más utilizado en bioestadística y en muchas otras disciplinas debido a su facilidad de interpetar y analizar.

El riesgo para un sujeto con un determinado conjunto de covariables nombradas $\{x_{1j}, x_{2j}, \ldots, x_{mj}\}$ dentro de este modelo se pueden descomponer en dos partes:

1. La primera involucra al tiempo, pero no a las covariables.
2. La segunda toma en cuenta las covariables, pero no el tiempo.

Dichas características se pueden observar en la siguiente fórmula,

$$h_j(t) = \exp\{\beta_1 x_{1j} + \ldots + \beta_m x_{mj}\} h_0(t)$$

Donde:

- $h_0(t)$ Representa la base para la función de riesgo.
- $\{\beta_1, \beta_2, \ldots, \beta_m\}$ Son las covariables que se toman como parámetros o coeficientes que se deben estimar.
- $\{t_1 < t_2 < \ldots < t_r\}$ Son los únicos tiempos ordenados de falla.

Cuando no hay relación entre la estimación de los coeficientes $\{\beta_1, \beta_2, \ldots, \beta_m\}$, se obtienen al incrementar al máximo la probabilidad parcial[5].

[5]Cox, pag. 35, (1972).

$$L(\beta_1, \beta_2, \ldots, \beta_m) = \prod_{k=1}^{r} \left(\frac{\exp\{\beta_1 x_{1k} + \ldots + \beta_m x_{mk}\}}{\sum_{j \in R_k} \exp\{\beta_1 x_{1j} + \ldots + \beta_m x_{mj}\}} \right)$$

Donde el producto asume todos los tiempos de falla t_k, y R_k representa el conjunto de individuos en riesgo al tiempo, t_k.

El modelo de Cox es "semiparamétrico", no tiene específicada la función de riesgo, esto permite estimar los coeficientes de la regresión, calcular las razones de riesgo y ajustar las curvas de supervivencia a diversas situaciones.

El modelo de riesgos proporcionales de Cox también es "robusto" en el sentido de que los resultados obtenidos en los ajustes tienden a aproximarse a los adquiridos por un modelo paramétrico. Esto quiere decir, que si el modelo correcto para el estudio es el exponencial (o lo mismo para el Weibull) las curvas de supervivencia obtenidas con el modelo de Cox son similares a las obtenidas con el modelo exponencial.

El modelo de Cox es una buena alternativa para modelos paramétricos[6]. El análisis de regresión clásico para riesgos competitivos establece un modelo de riesgos proporcionales de Cox para cada riesgo de causa-específica y se define como:

$$\lambda_j(t|Z) = \lambda_{0j} \mathrm{e}^{\beta_{\mathrm{j}}' \mathrm{Z}} \qquad \text{con} \qquad \mathrm{j} = 1, 2, \ldots, \mathrm{p}$$

Donde Z representa un vector de p covariables y β_j un vector de p coeficientes de regresión para cada falla. De forma independiente se estudia cada tipo de falla, a los sujetos que presentaron la falla debido a otras causas, se les toma como observaciones censuradas. Se asume que el efecto de las covariables actúan multiplicativamente con base en una función de riesgo desconocida λ_{0j}, como se realiza en un análisis de riesgos proporcionales clásico.

[6]Eva Boj del Val, *El modelo de regresión de Cox*, Departamento de Matemática Económica, Financiera y Actuarial, Facultad de Economía y Empresa, Universidad de Barcelona, Enero de 2014.

Para la estimación de los parámetros de la regresión β_j se utiliza el método de verosimilitudes parciales. Suponiendo que una muestra aleatoria censurada $\{\alpha_k, \delta_k, \delta_k, \omega_k\}$ con $k = 1, \ldots, r$, tiene r tiempos distintos de falla observados $\{t_1 < t_2 < \ldots t_R\}$ y considerando que r-R tiempos de censura (no considerados aquí). Tomando en cuenta la probabilidad de que un sujeto presente la falla debido a la causa j, al tiempo t_R, dado que uno de los individuos en riesgo (vivo y sin censura) presenta la falla al tiempo t_k, por la causa j, se expresa de la siguiente forma:

$$L(\beta_1, \ldots, \beta_p) = \prod_{k=1}^{r} \prod_{j=1}^{p} \left(\frac{e^{\beta_j' Z_k}}{\sum_{\mu=1}^{r} Y_\mu(t_k) e^{\beta_j' Z_\mu}} \right)^{\delta_{kj}} \qquad (52)$$

Donde $Y_\mu(t) = (t_\mu \geq t)$. La función parcial de verosimilitud se define sólo en los R tiempos de falla. Y $\delta_{kj} = (C = j)$, es el grupo de riesgo el cual puede disminuir por la ocurrencia de un evento debido a cualquier causa.

Si se maximiza cada uno de los factores de la ecuación anterior se obtiene un estimador $\widehat{\beta}_j$ consistente y asintóticamente normal bajo condiciones apropiadas. Dado $\widehat{\beta}_j$, el estimador de Nelson-Aalen[7] generalizado que estima las funciones de riesgo acumulado con bases relativas a cada causa-específica.

$$\widehat{\wedge}(t) = \sum_{k:t_\mu \leq t} \left(\frac{d_{kj}}{\sum_{\mu=1}^{r} Y_\mu(t_k) e^{\widehat{\beta}_j' Z_\mu}} \right) \qquad \text{con} \qquad j = 1, 2, \ldots, p$$

La inferencia para las $\widehat{\beta}_j$ y las $\widehat{\wedge}$ puede llevarse a cabo al igual que en los modelos estándares de Cox, donde se considera un sólo tipo de falla.

[7]El estimador de Nelson-Aalen es no paramétrico de la función de riesgo acumulado para el caso de datos censurados o datos incompletos.

La función de supervivencia y la de riesgo acumulado para t, dado Z, se obtiene por,

$$\widehat{S}(t|Z) = \exp\left\{-\sum_{j=1}^{p} \widehat{\wedge}_{0j}(t)\mathrm{e}^{\widehat{\beta}_{\mathrm{j}}'\mathrm{Z}_{\mu}}\right\} \qquad (53)$$

$$\widehat{\wedge}_{j}(t|Z) = \widehat{\wedge}_{0j}(t)\mathrm{e}^{\widehat{\beta}_{\mathrm{j}}'\mathrm{Z}_{\mu}} \qquad \text{con} \qquad \mathrm{j} = 1, 2, \ldots, \mathrm{p}$$

Finalmente, la función de incidencia acumulada $F_j(t|Z)$, también debe ser estimada y puede obtenerse de la siguiente forma:

$$\begin{aligned} \widehat{F}_j(t|Z) &= \int_0^t \widehat{S}(u|Z)d\widehat{\wedge}_j(u|Z) \\ &= \sum_{k:t_k \leq t} \delta_{kj} \exp\left\{-\sum_{\lambda=1}^{p} \widehat{\wedge}_{0\mu}(u)\right\} \frac{\mathrm{e}^{\widehat{\beta}_{\mathrm{j}}'\mathrm{Z}}}{\sum_{m=1}^{r} Y_r(t_k)\mathrm{e}^{\widehat{\beta}_{\mathrm{j}}'\mathrm{Z}_{\mathrm{m}}}} \end{aligned} \qquad (54)$$

Es muy importante modelar todos los tipos de falla con el fin de obtener una interpretación adecuada y completa del evento de interés.

Riesgos proporcionales de causa-específica

Al observar el comportamiento de $\log(-\log(S))$ vs. el tiempo de falla, se puede decir que los riesgos de causa-específica son proporcionales. En este caso S es el estimador de Kaplan-Meier cuando el único evento que se considera es el de interés; es decir, tanto las observaciones sin un evento y los riesgos competitivos se toman como censura. Se puede obtener para cada uno el nivel de las covariables. Si las dos curvas son paralelas la afirmación se cumple. Si la covariable es continua y se compara una parte del residuo de Schoenfeld contra el tiempo, indicará si hay una desviación.

Los denominados residuos de Schoenfeld se emplean para verificar el modelo de regresión de Cox. Siendo éstos los más efectivos en cuanto a detectar anomalías para cada una de las variables que intervienen en el modelo. Estos residuos pueden sugerir cambios, por ejemplo, que se utilice alguna

transformación para los datos. En el caso de los residuos de Schoenfeld se tiene uno para cada variable y paciente, es decir, si se tiene un modelo de Cox con tres factores pronóstico, se calcularán 3 residuos de Schoenfeld por paciente, valen cero cuando hay observaciones incompletas, para facilitar su interpretación se suelen presentar en las salidas sólo para los pacientes fallecidos.

En el modelo de riesgos de causa-específica así como en el de riesgo de la función de incidencia acumulada se supone que el logaritmo del riesgo cambia linealmente con la covariable. La afirmación de linealidad puede comprobarse por la categorización de la covariable y examinación de los coeficientes para cada categoría. Si los coeficientes aumentan linealmente la afirmación se satisface. Siempre hay que tomar en cuenta que el número de categorías es dictado por el tamaño del conjunto de datos y de manera más precisa por el número de eventos presentes.

Apéndice B. Códigos en R.3.1.1

Estimador de Kaplan-Meier

Ejemplo 1. Estimador de Kaplan-Meier, función de supervivencia.

```
gente<- Surv(T,S==1)
gente.all <-survfit(gente~1, conf.type=''plain'',
conf.int=.9)
gente.all
summary(gente.all)
plot(gente.all)

records    n.max n.start   events   median   0.9LCL   0.9UCL
   20       20       20       11       22       13       34

 time n.risk n.eve surv    std.err    lower90%CI  90%CI
    6     19     1   0.947   0.0512       0.8631     1.000
    8     16     1   0.888   0.0748       0.7651     1.000
   10     15     1   0.829   0.0902       0.6805     0.977
   11     14     1   0.770   0.1014       0.6030     0.936
   12     11     1   0.700   0.1138       0.5126     0.887
   13     10     2   0.560   0.1270       0.3510     0.769
   22      4     1   0.420   0.1541       0.1663     0.673
   32      3     1   0.280   0.1537       0.0271     0.533
   34      2     1   0.140   0.1253       0.0000     0.346
   36      1     1   0.000      NaN          NaN       NaN
```

Ejemplo 1. Estimador de Kaplan-Meier, función de riesgo acumulado.

```
plot(gente.all,conf.int=T, fun=``cumhaz'', lwd=2,
 col=``purple'',
 main="Funcion de Riesgo Acumulado
+ Osteosarcoma", xlab="Tiempo(meses)")
box(lwd=3, col = 'black')
legend("bottomright",c("Riesgo Acumulado"),lty=1,
 col="purple")
```

Ejemplo 2. Estimador de Kaplan-Meier.

```
gente<- Surv(T,S==1)
gente.all <-survfit(gente~1, conf.type="plain",
conf.int=.9)
gente.all
Call: survfit(formula = gente ~ 1, conf.type = "plain",
conf.int = 0.9)

records   n.max n.start   events   median   0.9LCL   0.9UCL
  30        30     30        17       32       24       39
summary(gente.all)
Call: survfit(formula = gente ~ 1, conf.type = "plain",
 conf.int = 0.9)

 time n.risk n.event   surv   td.err   90%CI   90% CI
  2      29       1    0.9655   0.0339   0.9098   1.000
  3      28       1    0.9310   0.0471   0.8536   1.000
  5      26       1    0.8952   0.0573   0.8010   0.989
  8      25       1    0.8594   0.0652   0.7521   0.967
  9      24       1    0.8236   0.0717   0.7057   0.941
 13      21       1    0.7844   0.0783   0.6557   0.913
 15      19       1    0.7431   0.0843   0.6044   0.882
 19      17       1    0.6994   0.0900   0.5514   0.847
 24      14       1    0.6494   0.0964   0.4908   0.808
 28      12       1    0.5953   0.1025   0.4268   0.764
 29      11       1    0.5412   0.1065   0.3660   0.716
```

32	10	1	0.4871	0.1087	0.3082	0.666
34	9	1	0.4330	0.1093	0.2532	0.613
35	8	1	0.3788	0.1082	0.2009	0.557
39	4	1	0.2841	0.1154	0.0943	0.474
48	3	1	0.1894	0.1091	0.0100	0.369
50	2	1	0.0947	0.0864	0.0000	0.237

Capítulo III. Análisis de datos: cáncer de cuello uterino en fase primaria.

```
#Variable HP5
ccup<-read.csv("C:/Users/Documents/ccup_rc.csv",
header=T)
attach(ccup)
names(ccup)
 [1] ''edad'' ''hgb'' ''tamtum'' ''IFP'' ''HP5''
     ''pelvi''
 [7] ''resp'' ''pelrec'' ''disrec'' ''survtime''
     ''stat'' ''dftime''
[13] ''dfcens''

edad
  [1] 78 69 55 55 50 57 53 62 23 57 74 67 72 66 47 61
      78 52 33 45 38 55 31 70 75
 [26] 71 47 67 60 41 27 55 38 30 43 48 36 35 66 72 63
      68 66 54 59 72 32 43 74 35
 [51] 41 46 65 46 57 27 45 33 49 61 33 67 66 55 43 77
      53 44 68 51 49 23 32 64 71
 [76] 41 54 46 62 42 49 61 48 68 74 40 65 51 58 43 53
      47 74 49 46 51 47 42 44 54
[101] 28 45 61 66 65 43 47 46 77

hgb
  [1] 119 131 126 141  95 132 127 142 145 142 124 133
      133 116  82 118  95 150
```

```
 [19] 119 125 127 143 105 125 151 142 114 145 120 132
      133 123 100 137 112 124
 [37] 115 116  99 124 129 135 136 147 124 123 143 117
      130 109 109 126 139 129
 [55] 152 124 129 122 139 139 151 123 121 119 105 140
      132 122 128 114 123 136
 [73]  95 129 138 137 124 144 106 124  84 123 109 138
      107 130 124 123 132 117
 [91] 134 137 122 121 136 140 141 130 124 124 103 128
      128 132 133 107 136 127
[109] 121
```

tamtum

```
  [1]  7.0  2.0 10.0  8.0  8.0  8.0  4.0  5.0  5.0  3.0
       4.0  5.0  4.0  8.0 10.0
 [16]  5.0  7.0  6.0  7.0  5.0  9.0 10.0  5.0  6.0  3.0
       5.0  5.0  4.0  8.0  5.0
 [31]  5.0  4.0  5.0  4.0  5.0  4.0  8.0  6.0  7.0  6.0
       3.5  6.0  5.0  8.0  5.0
 [46]  6.0  3.0  6.0  4.5  6.0  4.5  8.0  5.0  4.0  3.0
       5.0  3.0  5.0  6.0  3.0
 [61]  5.0  4.0  4.0  4.0  5.0  4.0  6.0  8.0  6.0  8.0
       5.0  4.0  7.5  4.0  3.5
 [76]  8.0  5.0  5.0 10.0  8.0 10.0  4.0  7.0  7.0  7.0
       5.0  5.0  5.0  5.0  7.0
 [91]  3.0  4.0  5.0  8.0  7.0  5.0  4.0  8.0  7.0  8.0
       5.0  4.0  4.0  3.0  5.0
[106]  6.0  5.0  4.0  5.0
```

IFP

```
  [1]  8.0  8.2  8.6  3.3 18.5 20.0 21.8 31.6 16.5 31.5
      18.5 12.8 18.4 18.5 21.0
 [16] 23.6 21.0 11.1 14.6 30.9 19.6 24.0 15.8 15.0 13.2
      16.8 16.8  6.6 24.8 12.0
 [31] 37.4 10.3 19.4 14.2 16.6 13.9 47.9 14.5 29.9 17.5
      23.6 18.5 24.8 34.6  7.1
 [46]  6.5 19.6 32.0 12.9 16.3 13.8 22.3 12.0  9.8 19.4
```

```
      18.7  23.1  12.1  23.4  12.5
 [61] 11.7  12.9  11.2  23.6  26.7  21.5  28.9  21.1  23.2  43.5
      33.3  10.0  11.0  42.1  14.0
 [76] 18.5  18.5  28.5  20.2  20.6  22.7  38.0  20.9   2.8  27.6
      25.6  38.4   9.5  18.1  28.7
 [91] 19.8  23.2  19.5  46.2  39.7  18.9   8.5  22.3  26.5  14.3
      16.9   9.6   9.4  23.6  21.2
[106] 14.6  12.2  12.9  16.7
```

HP5

```
  [1] 32.1428571  2.1739130 52.3255814  3.2608696
      85.4304636 19.3548387
  [7] 44.5783133 59.6774194 29.1666667 85.7142857
       8.0645161 77.6315789
 [13] 33.3333333 99.2187500 66.2921348 55.0000000
      81.6000000 56.7567568
 [19] 52.0408163 46.7741935 43.7500000 91.7293233
      12.7450980 32.5000000
 [25] 82.5000000 57.7777778 70.6293706 12.2807018
      47.0198676 17.1641791
 [31] 34.3750000 64.8648649 35.7142857 58.8888889
      12.5984252 30.2325581
 [37] 54.8872180 63.4782609 87.3873874 26.0162602
      24.8000000 63.0136986
 [43] 67.5000000 27.6315789 44.8275862 28.1609195
      51.8750000 78.8079470
 [49] 11.2500000 11.8750000 50.0000000 88.7500000
      27.3437500 31.5436242
 [55] 64.9253731 29.0502793 43.6241611 30.6250000
      80.8917197 58.0246914
 [61] 58.7500000 19.5652174 19.5652174 77.2972973
      60.7142857 35.5555556
 [67] 81.7610063 65.6250000 87.7906977 53.1645570
      78.2608696 62.7329193
 [73] 93.0817610 58.7500000 41.3043478 66.1870504
      83.0357143 73.7500000
 [79] 71.8750000 59.1397849 18.7500000 17.8947368
```

```
      55.6250000   4.3478261
 [85] 92.5000000   0.6622517   0.0000000 69.3333333
       9.8360656 59.0551181
 [91] 14.9532710   0.0000000   7.6271186 44.9367089
       2.3121387 12.3287671
 [97] 28.8461538 20.6896552 11.8055556 88.6792453
      44.6540880 82.7160494
[103] 28.1481481   1.8750000 51.3043478 49.1891892
      89.3939394 51.8181818
[109] 26.8292683

pelvi
  [1] N N N N Y N E N N N N Y N E Y N N N N N E N Y N
      Y N E N Y Y N N E Y N N N
 [38] N N N N N E E N E Y N N E E N N N N E N N N N N
      N N N E N E Y Y Y E E E E
 [75] N Y N E Y N Y N Y Y Y N E Y N N E N N N N N E Y
      N N E Y N N N N Y E N
Levels: E N Y

resp
  [1] CR CR NR CR NR CR CR CR CR CR CR NR CR NR CR CR
      NR CR CR CR CR CR NR CR CR
 [26] CR CR CR NR CR NR CR CR CR CR CR NR CR CR CR CR
      CR CR CR CR CR CR CR CR CR
 [51] CR NR CR CR CR CR CR CR CR CR CR CR CR CR CR CR
      CR NR CR NR CR CR NR CR CR
 [76] CR CR CR CR CR NR CR NR CR CR CR CR CR CR CR CR
      CR CR NR CR CR CR NR CR CR
[101] CR CR CR CR CR CR CR CR CR
Levels: CR NR

pelrec
  [1] N N Y Y Y N N Y N N N Y Y Y N N Y N Y N N N Y N
      N N Y N Y N Y N N Y N N Y
 [38] N N N N Y N Y N Y Y N N N N Y N N N N Y N N N N
      N N N N N N Y N Y N N Y N
```

```
 [75] N N Y N Y N Y N Y N N Y N N N N N N N Y Y N N Y
      N N Y N N N N N N N N
Levels: N Y

disrec
  [1] N N N Y N N N Y N N N Y N Y Y Y N N Y N N Y N N
      N N Y N N Y Y N N N N Y Y
 [38] N N N N N Y Y N N N Y N N N Y N N N N Y N Y N N
      Y N Y Y N Y Y N N Y N N Y
 [75] N N N N Y N Y Y Y N N N N Y N N Y N N N N Y N N
      N N N N N N N N N N N N

Levels: N Y

survtime
  [1] 6.152 8.008 0.621 1.120 1.292 7.929 8.454 7.116
      8.378 8.178 3.395 1.016
 [13] 3.699 0.630 8.194 4.764 2.590 7.707 1.478 7.316
      7.841 1.133 1.268 7.543
 [25] 6.300 7.587 3.121 6.957 0.841 1.766 1.344 6.114
      6.374 7.277 5.714 2.779
 [37] 1.098 6.812 1.689 6.097 5.421 2.938 6.108 2.294
      5.988 2.278 1.949 3.784
 [49] 2.779 6.442 2.100 1.287 6.360 0.901 6.272 3.316
      2.839 5.938 1.514 5.199
 [61] 5.306 4.331 3.127 2.196 2.628 5.153 2.313 1.210
      1.076 5.791 4.482 4.446
 [73] 0.980 1.106 3.348 5.380 5.350 4.786 2.593 5.021
      0.775 5.276 2.094 4.405
 [85] 2.875 5.005 4.433 2.108 4.526 4.057 2.886 2.916
      3.559 0.663 2.168 2.812
 [97] 4.183 1.684 4.408 4.016 1.927 4.112 3.833 4.153
      3.775 3.784 3.901 3.606
[109] 3.288

stat
  [1] 0 0 1 1 1 0 0 0 0 0 0 1 1 1 0 1 0 0 1 0 0 1 1 0
      0 0 1 0 1 1 1 0 0 0 0 1 1
```

```
 [38] 0 0 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 0 0 1 0 1 0 0
      0 0 1 1 0 1 1 0 0 0 0 1 1
 [75] 0 0 0 0 1 0 1 0 1 0 0 0 0 1 0 0 1 0 0 1 1 1 0 0
      0 0 1 0 0 0 0 0 0 0 0

dftime
  [1] 6.152 8.008 0.003 1.073 0.003 7.929 8.454 7.107
      8.378 8.178 3.395 0.003
 [13] 1.350 0.003 0.512 1.714 0.003 7.707 0.939 7.316
      7.841 0.589 0.003 7.543
 [25] 6.300 7.587 1.123 6.957 0.003 0.460 0.003 6.114
      6.374 2.519 5.714 1.514
 [37] 0.003 6.812 1.689 6.097 5.421 1.246 6.108 1.084
      5.988 1.530 1.029 1.068
 [49] 2.779 6.442 2.100 0.003 6.360 0.901 6.272 3.316
      2.606 5.938 1.013 5.199
 [61] 5.306 3.814 3.127 0.709 1.610 5.153 1.391 0.000
      1.076 0.003 1.785 4.446
 [73] 0.003 0.690 3.348 5.380 0.991 4.786 1.541 5.021
      0.003 4.988 0.003 4.405
 [85] 2.875 1.741 4.433 0.868 4.526 4.057 1.941 2.916
      3.559 0.003 1.112 1.873
 [97] 4.183 0.003 4.408 4.016 0.706 4.112 3.833 4.153
      3.775 3.784 3.901 3.606
[109] 3.288

dfcens
  [1] 0 0 1 1 1 0 0 1 0 0 0 1 1 1 1 1 1 0 1 0 0 1 1 0
      0 0 1 0 1 1 1 0 0 1 0 1 1
 [38] 0 0 0 0 1 0 1 0 1 1 1 0 0 0 1 0 0 0 0 1 0 1 0 0
      1 0 1 1 0 1 1 0 1 1 0 1 1
 [75] 0 0 1 0 1 0 1 1 1 0 0 1 0 1 0 0 1 0 0 1 1 1 0 1
      0 0 1 0 0 0 0 0 0 0 0

#Matriz de covariables.
matricov
             [,1] [,2] [,3]
```

```
      [1,] 32.1428571  7.0    1
      [2,]  2.1739130  2.0    1
      [3,] 52.3255814 10.0    1
      [4,]  3.2608696  8.0    1
      [5,] 85.4304636  8.0    0
      [6,] 19.3548387  8.0    1
      [7,] 44.5783133  4.0    1
      [8,] 59.6774194  5.0    1
      [9,] 29.1666667  5.0    1
     [10,] 85.7142857  3.0    1
     [11,]  8.0645161  4.0    1
     [12,] 77.6315789  5.0    0
     [13,] 33.3333333  4.0    1
     [14,] 99.2187500  8.0    1
     [15,] 66.2921348 10.0    0
     [16,] 55.0000000  5.0    1
     [17,] 81.6000000  7.0    1
     [18,] 56.7567568  6.0    1
     [19,] 52.0408163  7.0    1
     [20,] 46.7741935  5.0    1
     [21,] 43.7500000  9.0    1
     [22,] 91.7293233 10.0    1
     [23,] 12.7450980  5.0    0
     [24,] 32.5000000  6.0    1
     [25,] 82.5000000  3.0    0
     [26,] 57.7777778  5.0    1
     [27,] 70.6293706  5.0    1
     [28,] 12.2807018  4.0    1
     [29,] 47.0198676  8.0    0
     [30,] 17.1641791  5.0    0
     [31,] 34.3750000  5.0    1
     [32,] 64.8648649  4.0    1
     [33,] 35.7142857  5.0    1
     [34,] 58.8888889  4.0    0
     [35,] 12.5984252  5.0    1
     [36,] 30.2325581  4.0    1
     [37,] 54.8872180  8.0    1
```

[38,]	63.4782609	6.0	1
[39,]	87.3873874	7.0	1
[40,]	26.0162602	6.0	1
[41,]	24.8000000	3.5	1
[42,]	63.0136986	6.0	1
[43,]	67.5000000	5.0	1
[44,]	27.6315789	8.0	1
[45,]	44.8275862	5.0	1
[46,]	28.1609195	6.0	1
[47,]	51.8750000	3.0	0
[48,]	78.8079470	6.0	1
[49,]	11.2500000	4.5	1
[50,]	11.8750000	6.0	1
[51,]	50.0000000	4.5	1
[52,]	88.7500000	8.0	1
[53,]	27.3437500	5.0	1
[54,]	31.5436242	4.0	1
[55,]	64.9253731	3.0	1
[56,]	29.0502793	5.0	1
[57,]	43.6241611	3.0	1
[58,]	30.6250000	5.0	1
[59,]	80.8917197	6.0	1
[60,]	58.0246914	3.0	1
[61,]	58.7500000	5.0	1
[62,]	19.5652174	4.0	1
[63,]	19.5652174	4.0	1
[64,]	77.2972973	4.0	1
[65,]	60.7142857	5.0	1
[66,]	35.5555556	4.0	1
[67,]	81.7610063	6.0	1
[68,]	65.6250000	8.0	0
[69,]	87.7906977	6.0	0
[70,]	53.1645570	8.0	0
[71,]	78.2608696	5.0	1
[72,]	62.7329193	4.0	1
[73,]	93.0817610	7.5	1
[74,]	58.7500000	4.0	1

```
 [75,] 41.3043478  3.5    1
 [76,] 66.1870504  8.0    0
 [77,] 83.0357143  5.0    1
 [78,] 73.7500000  5.0    1
 [79,] 71.8750000 10.0    0
 [80,] 59.1397849  8.0    1
 [81,] 18.7500000 10.0    0
 [82,] 17.8947368  4.0    1
 [83,] 55.6250000  7.0    0
 [84,]  4.3478261  7.0    0
 [85,] 92.5000000  7.0    0
 [86,]  0.6622517  5.0    1
 [87,]  0.0000000  5.0    1
 [88,] 69.3333333  5.0    0
 [89,]  9.8360656  5.0    1
 [90,] 59.0551181  7.0    1
 [91,] 14.9532710  3.0    1
 [92,]  0.0000000  4.0    1
 [93,]  7.6271186  5.0    1
 [94,] 44.9367089  8.0    1
 [95,]  2.3121387  7.0    1
 [96,] 12.3287671  5.0    1
 [97,] 28.8461538  4.0    1
 [98,] 20.6896552  8.0    0
 [99,] 11.8055556  7.0    1
[100,] 88.6792453  8.0    1
[101,] 44.6540880  5.0    1
[102,] 82.7160494  4.0    0
[103,] 28.1481481  4.0    1
[104,]  1.8750000  3.0    1
[105,] 51.3043478  5.0    1
[106,] 49.1891892  6.0    1
[107,] 89.3939394  5.0    0
[108,] 51.8181818  4.0    1
[109,] 26.8292683  5.0    1

#predict.crr para HP5.
```

```
pfit
        [,1]        [,2]        [,3]        [,4]
 [1,] 0.460 0.004921683 0.003576141 0.009413754
 [2,] 0.512 0.009840464 0.007155017 0.018779397
 [3,] 0.589 0.014761787 0.010740619 0.028107275
 [4,] 0.690 0.019714692 0.014354187 0.037451974
 [5,] 0.709 0.024708523 0.018002680 0.046830153
 [6,] 0.868 0.029767622 0.021704080 0.056286106
 [7,] 1.013 0.034909152 0.025471208 0.065849913
 [8,] 1.068 0.040105653 0.029284202 0.075468612
 [9,] 1.391 0.045440499 0.033204594 0.085293834
[10,] 1.514 0.050836477 0.037176020 0.095180550
[11,] 1.610 0.056253008 0.041168796 0.105053220
[12,] 1.714 0.061833038 0.045288664 0.115169702
[13,] 1.785 0.067453565 0.049445222 0.125303975
[14,] 1.873 0.073149585 0.053664615 0.135517379
[15,] 1.941 0.078849926 0.057894326 0.145681070
[16,] 3.814 0.086043115 0.063241991 0.158424417
[17,] 4.988 0.096095211 0.070734423 0.176079032
[18,] 6.108 0.110535159 0.081537467 0.201126554

             [,5]
[1,] 0.006844353
[2,] 0.013671427
[3,] 0.020488735
[4,] 0.027336329
[5,] 0.034226787
[6,] 0.041193213
[7,] 0.048258573
[8,] 0.055384495
[9,] 0.062684413
[10,] 0.070051731
[11,] 0.077430616
[12,] 0.085014892
[13,] 0.092636356
[14,] 0.100341839
[15,] 0.108034607
```

```
[16,] 0.117715405
[17,] 0.131193816
[18,] 0.150452996

attr(,"class")
[1] "predict.crr"


#Plot.predic.crr
par(las=1,mfrow=c(1,1))
plot.predict.crr(pfit,lty=c(1:4),ylim=c(0,1),color=1:4,
xlab="Recada local (años)", col.main="black"
ylab="Función de Incidencia Acumulada",
main="Estimación de los grupos de la Función de
Incidencia Acumulada de HP5",)
legend(0,1,lty=c(1:4),bty="n",
legend=c("HP5=5, Tamaño del tumor=5,
Nodo negativo/equivoco.",
"HP5=5, Tamaño del tumor=5, Nodo positivo.",
"HP5=46, Tamaño del tumor=5, Nodo negativo/equivoco.",
"HP5=46, Tamaño del tumor=5, Nodo positivo."))

############################################################

#Variable IFP

#Influencia de la variable IFP en el evento de interes:
recaida local.

#Covariables para la variable.

matripelve
      [,1] [,2] [,3]
 [1,]  8.0  7.0    1
 [2,]  8.2  2.0    1
 [3,]  8.6 10.0    1
 [4,]  3.3  8.0    1
```

```
 [5 ,]  18.5   8.0     0
 [6 ,]  20.0   8.0     1
 [7 ,]  21.8   4.0     1
 [8 ,]  31.6   5.0     1
 [9 ,]  16.5   5.0     1
[10 ,]  31.5   3.0     1
[11 ,]  18.5   4.0     1
[12 ,]  12.8   5.0     0
[13 ,]  18.4   4.0     1
[14 ,]  18.5   8.0     1
[15 ,]  21.0  10.0     0
[16 ,]  23.6   5.0     1
[17 ,]  21.0   7.0     1
[18 ,]  11.1   6.0     1
[19 ,]  14.6   7.0     1
[20 ,]  30.9   5.0     1
[21 ,]  19.6   9.0     1
[22 ,]  24.0  10.0     1
[23 ,]  15.8   5.0     0
[24 ,]  15.0   6.0     1
[25 ,]  13.2   3.0     0
[26 ,]  16.8   5.0     1
[27 ,]  16.8   5.0     1
[28 ,]   6.6   4.0     1
[29 ,]  24.8   8.0     0
[30 ,]  12.0   5.0     0
[31 ,]  37.4   5.0     1
[32 ,]  10.3   4.0     1
[33 ,]  19.4   5.0     1
[34 ,]  14.2   4.0     0
[35 ,]  16.6   5.0     1
[36 ,]  13.9   4.0     1
[37 ,]  47.9   8.0     1
[38 ,]  14.5   6.0     1
[39 ,]  29.9   7.0     1
[40 ,]  17.5   6.0     1
[41 ,]  23.6   3.5     1
```

```
[42,] 18.5  6.0   1
[43,] 24.8  5.0   1
[44,] 34.6  8.0   1
[45,]  7.1  5.0   1
[46,]  6.5  6.0   1
[47,] 19.6  3.0   0
[48,] 32.0  6.0   1
[49,] 12.9  4.5   1
[50,] 16.3  6.0   1
[51,] 13.8  4.5   1
[52,] 22.3  8.0   1
[53,] 12.0  5.0   1
[54,]  9.8  4.0   1
[55,] 19.4  3.0   1
[56,] 18.7  5.0   1
[57,] 23.1  3.0   1
[58,] 12.1  5.0   1
[59,] 23.4  6.0   1
[60,] 12.5  3.0   1
[61,] 11.7  5.0   1
[62,] 12.9  4.0   1
[63,] 11.2  4.0   1
[64,] 23.6  4.0   1
[65,] 26.7  5.0   1
[66,] 21.5  4.0   1
[67,] 28.9  6.0   1
[68,] 21.1  8.0   0
[69,] 23.2  6.0   0
[70,] 43.5  8.0   0
[71,] 33.3  5.0   1
[72,] 10.0  4.0   1
[73,] 11.0  7.5   1
[74,] 42.1  4.0   1
[75,] 14.0  3.5   1
[76,] 18.5  8.0   0
[77,] 18.5  5.0   1
[78,] 28.5  5.0   1
```

```
 [79,] 20.2 10.0    0
 [80,] 20.6  8.0    1
 [81,] 22.7 10.0    0
 [82,] 38.0  4.0    1
 [83,] 20.9  7.0    0
 [84,]  2.8  7.0    0
 [85,] 27.6  7.0    0
 [86,] 25.6  5.0    1
 [87,] 38.4  5.0    1
 [88,]  9.5  5.0    0
 [89,] 18.1  5.0    1
 [90,] 28.7  7.0    1
 [91,] 19.8  3.0    1
 [92,] 23.2  4.0    1
 [93,] 19.5  5.0    1
 [94,] 46.2  8.0    1
 [95,] 39.7  7.0    1
 [96,] 18.9  5.0    1
 [97,]  8.5  4.0    1
 [98,] 22.3  8.0    0
 [99,] 26.5  7.0    1
[100,] 14.3  8.0    1
[101,] 16.9  5.0    1
[102,]  9.6  4.0    0
[103,]  9.4  4.0    1
[104,] 23.6  3.0    1
[105,] 21.2  5.0    1
[106,] 14.6  6.0    1
[107,] 12.2  5.0    0
[108,] 12.9  4.0    1
[109,] 16.7  5.0    1

#FIC construida con el riesgo acumulado de
la subdistribuci n
# pfit=predic.crr a los escenarios de IFP.

pfit
```

```
      [,1]        [,2]        [,3]        [,4]
 [1,] 0.000 0.003751291 0.00501647 0.01023908
 [2,] 0.003 0.030007740 0.03994871 0.08004579
 [3,] 0.460 0.034164712 0.04545022 0.09080184
 [4,] 0.512 0.038335581 0.05096207 0.10151328
 [5,] 0.589 0.042586465 0.05657136 0.11234742
 [6,] 0.690 0.046904518 0.06226066 0.12326746
 [7,] 0.709 0.051278979 0.06801540 0.13424281
 [8,] 0.868 0.055670146 0.07378311 0.14517195
 [9,] 0.939 0.060105294 0.07959938 0.15612123
[10,] 1.013 0.064561023 0.08543331 0.16703125
[11,] 1.068 0.069047383 0.09129786 0.17792537
[12,] 1.073 0.073587150 0.09722250 0.18885667
[13,] 1.084 0.078223450 0.10326300 0.19992471
[14,] 1.123 0.082958250 0.10942124 0.21112845
[15,] 1.391 0.087707978 0.11558810 0.22226693
[16,] 1.514 0.092508621 0.12181005 0.23342284
[17,] 1.541 0.097315231 0.12802859 0.24449029
[18,] 1.610 0.102231007 0.13437679 0.25570364
[19,] 1.714 0.107309823 0.14092321 0.26717733
[20,] 1.785 0.112421089 0.14749876 0.27861032
[21,] 1.873 0.117595764 0.15414285 0.29006914
[22,] 1.941 0.122793324 0.16080307 0.30146164
[23,] 2.606 0.128071491 0.16755294 0.31291136
[24,] 3.814 0.134683745 0.17598941 0.32708599
[25,] 4.988 0.144157251 0.18803853 0.34706855
[26,] 6.108 0.158472296 0.20615998 0.37654190
[27,] 7.107 0.182442253 0.23627029 0.42397624
```

```
           [,5]
[1,] 0.009162013
[2,] 0.071898881
[3,] 0.081609477
[4,] 0.091291860
[5,] 0.101097577
[6,] 0.110993854
```

```
[7,] 0.120953381
[8,] 0.130884244
[9,] 0.140846877
[10,] 0.150787380
[11,] 0.160727139
[12,] 0.170714817
[13,] 0.180841938
[14,] 0.191108310
[15,] 0.201330171
[16,] 0.211583541
[17,] 0.221771213
[18,] 0.232109276
[19,] 0.242704412
[20,] 0.253279422
[21,] 0.263896105
[22,] 0.274469313
[23,] 0.285113996
[24,] 0.298318036
[25,] 0.316982466
[26,] 0.344622610
[27,] 0.389400606
attr(,"class")
[1] "predict.crr"
```

Simulación

```
Competing Risks Regression

Call:
crr(ftime = ftime, fstatus = fstatus, cov1 = cov)

       coef exp(coef) se(coef)       z p-value
x1  -0.2738     0.760    0.416 -0.6577    0.51
x2  -0.0141     0.986    0.437 -0.0323    0.97
x3  -0.3781     0.685    0.443 -0.8526    0.39

   exp(coef) exp(-coef)  2.5% 97.5%
x1     0.760       1.31 0.336  1.72
x2     0.986       1.01 0.419  2.32
x3     0.685       1.46 0.287  1.63

Num. cases = 200
Pseudo Log-likelihood = -289
Pseudo likelihood ratio test = 1.26  on 3 df,
```

Bibliografía

- Aalen, O. (1993). Further results on the non-parametric linear regression model in survival analysis.
- Cox, D.R. (1959). The analysis of exponentially distributed lifetime with two types of failure. Journal of the Royal Statistical Society.
- Collett, D. (1994). Modelling Survival Data in Medical Research. Chapman and Hall.
- Crowder, M.J. (2001). Classical competing risks. Chapman Hall/CRC: Boca raton.
- Fernández Pita S. (2004). Análisis de supervivencia. Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario-Universitario Juan Canalejo. A Coruña (España).
- Fine, J. y Gray, R. (1999). A proportional hazards model for the subdistribution of a competing risk. Journal of the American Statistical Association.
- Godoy Aguilar, A.M. (2009). Introducción al Análisis de Supervivencia con R.
- Haesook, T. K. (2007). Cumulative Incidence in Competing Risks Data and Competing Risks Regression Analysis. American Journal of the Association for Cancer Research.
- Kalbfleisch J.D. y Prentice R.L. (2002). The Statistical Analysis of Failure Time Data. Wiley Series in Probability and Statistics.

- Lindqvist H. (2006). A review of Competing Risks, Departament of Mathematical Sciences, Norwegian University of Science and Technology.
- Matadamas Segura M.A. (2010). Inferencia para modelos de supervivencia de un solo evento y extensiones para modelos de riesgos competitivos. Universidad Autónoma Metropolitana. Unidad-Iztapalapa.
- Moeschberger M.L. and Klein J.P. (1995). Statistical methods for dependent competing risks. Lifetime Data Analysis.
- Package "cmprsk". (2014).
- Paz M.C, Canal Yañez S. (2010). Comparación de dos Grupos en Presencia de Riesgos Competitivos. Universidad Nacional de Colombia, Sede Medellín.
- Pintilie M. (2006). Competing risks: a practical perspective.
- Rodríguez Muñoz J.E. (2011).Modelación de datos espaciales censurados. Universidad de Guanajuato.
- Alba Gutiérrez, S. (2010). Modelos de riesgos competitivos. Proyecto Fin de Máster. Universidad de Granada.
- Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. Proceedings of National Academy of Sciences USA.