UNIVERSIDAD DE SEVILLA

Modelo de regresión de Cox y sus aplicaciones biosanitarias

Trabajo Fin de Grado - Grado en Estadística

Departamento de Estadística e I.O.

Facultad de Matemáticas

Curso Académico 2015/16

TRABAJO REALIZADO POR:

PAOLA VELASCO ÁLVAREZ

TUTOR:

JUAN M. MUÑOZ PICHARDO

Modelo de regresión de Cox y sus aplicaciones biosanitarias

Paola Velasco Álvarez

Trabajo fin de Grado

Grado en Estadística

Universidad de Sevilla

TUTOR

D. Juan M. Muñoz Pichardo

Departamento de Estadística e Investigación Operativa

Abstract

This project is called “Cox Proportional Hazards Model and its Life Sciences Applications”.

The aim of this project is to explain in detail the "Cox Proportional Hazard Model", whose model is a semiparametric survival one.

In chapter I, you will find a brief introduction about survival analysis, and some examples where this specific type may be used.

In chapter II, Cox’s regression model with its statistical inference, i.e., parameter estimation, hyphothesis contrast, adjustment and diagnostic analysis of the model are described.

In chapter III, some examples about Cox’s regression model are described and to conclude a complete example is carried out in R-Program.

Índice

Capítulo I. Introducción al Análisis de Supervivencia.

En este capítulo se incluye una breve introducción al análisis de supervivencia con los conceptos básicos, los tipos de censura de datos y las funciones que se permiten describir los fenómenos reales analizados por este conjunto de técnicas estadísticas. Por otro lado, se introducen los métodos no paramétricos y los métodos paramétricos. En los métodos no paramétricos se incluye una descripción de las estimaciones mediante Kaplan-Meier y la comparación de curvas de supervivencia; y en los métodos paramétricos se recogen las distribuciones más usadas, así como una breve presentación de los modelos de regresión más utilizados: modelo de tiempo de fallo acelerado y modelo de riesgos proporcionales de Cox. Y por último, se incluye un apartado con aplicaciones ilustrativas generales del análisis de supervivencia.

I.1. Conceptos básicos del Análisis de Supervivencia.

El análisis de supervivencia es una técnica inferencial cuyo principal objetivo es examinar y modelar el tiempo que se toma para que ocurra un determinado suceso. Generalmente, dada las innumerables aplicaciones de estas técnicas en el ámbito biomédico, el suceso es comúnmente denominado “muerte” y el tiempo como “tiempo de vida”

Este análisis consiste en un seguimiento continuo de una serie de individuos desde que comienza el estudio hasta que finaliza.

En este estudio, la variable de interés es conocida como “tiempo de vida” o “tiempo de supervivencia”; además, tenemos que tener en cuenta las “observaciones censuradas”, que son aquellas observaciones que desaparecen del estudio o aquellas observaciones en las que no se produce el suceso de interés antes de finalizar el estudio. Las observaciones censuradas nos dan información parcial sobre la variable tiempo bajo estudio.

Dichas censuras pueden presentarse de varias formas:

- ***Censura tipo I***: Observamos un número de sujetos desde el instante $t = 0$ incluyendo el tiempo en el que falla cada uno y finaliza en un tiempo fijado previamente $t = t_f$. Los sujetos que al finalizar el estudio no presenten fallo, formaran las observaciones censuradas. Si T denota el tiempo de fallo:

$$\begin{cases} T \leq t_f & Observado \\ T > t_f & Censurado \end{cases}$$

- ***Censura tipo II***: Esta censura es muy similar a la censura del tipo I pero en este caso el tiempo no está prefijado por el investigador sino que el experimento continua hasta que una fracción prefijada r/n (r fallos de n posibles), por lo que tenemos $n - r$ observaciones censuradas cuyos tiempo de fallos son desconocidos.

- ***Censura aleatoria***: Este tipo ocurre sin ningún control del investigador ya que pueden ser por diversas casusas. Por ejemplo, en un estudio biomédico:
  - Abandono: el paciente abandona el ensayo.
  - Salida forzosa: los efectos del tratamiento le obligan a dejar el estudio.
  - Fin del ensayo: termina el estudio y no se produce el suceso de interés.

- ***Censura por la derecha***: en el caso de presentarse el fallo, este se presenta después del tiempo de censura observado.

- ***Censura por la izquierda***: el sujeto presenta el fallo antes de ingresar en el estudio, por lo que su tiempo de fallo no observado es menor que el tiempo de censura observado.

- ***Censura por intervalo***: la observación de los sujetos no sucede de forma continua por lo que entre un tiempo de fallo y otro hay un periodo largo de observación.

Sea T el tiempo de supervivencia, consideramos T como una variable aleatoria continua. Denotaremos $f(x)$ como su función de densidad y $F(x)$ como su función de distribución.

Por ejemplo, en la distribución Exponencial dichas funciones son:

$$f(x) = \lambda e^{-\lambda x} \qquad\qquad F(x) = 1 - e^{-\lambda x}$$

Las siguientes funciones permiten dar respuesta en muchas de las preguntas que surgen en este tipo de estudios.

- ***Función de supervivencia***: es la probabilidad de que el sujeto estudiado sobreviva más de un periodo dado, t.

$$S(t) = \Pr[T > t] = 1 - F(x) = \int_t^{\infty} f(x)\, dx$$

- ***Función de riesgo o tasa de fallo instantánea***: evalúa el riesgo inmediato de muerte en el tiempo t condicionada a la supervivencia del individuo al instante t. Esta función puede tomar cualquier valor no negativo.

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr[t < T \leq t + \Delta t \,|T > t]}{\Delta t} = \frac{f(t)}{S(t)}$$

- ***Tasa de fallo acumulada***: se define integrando la función de riesgo.

$$H(t) = \int_0^t h(x)dx \quad ; \quad h(t) = \frac{d}{dt}H(t) \quad ; \quad H(t) = -\ln[S(t)]$$

- ***Tiempo esperado de vida***: también conocido como esperanza matemática de T.

$$E[T] = \int_0^\infty t\, f(t)\, dt = \int_0^\infty S(t)\; dt$$

- ***Vida media residual***: Esta función mide la esperanza de vida restante en los individuos de edad t ó el tiempo esperado de vida después de t, hasta que ocurre el fallo.

$$vmr(t) = E[T - t|\, T > t] = \frac{1}{S(t)} \int_t^\infty S(x)dx$$

I.2. Métodos no paramétricos.

Los métodos incluidos con el análisis de supervivencia se pueden clasificar en métodos paramétricos y métodos no paramétricos.

Entre los métodos no paramétricos, los más usados son el estimador de Kaplan-Meier y la comparación de curvas de supervivencia.

El principal objetivo del estimador de Kaplan-Meier es estimar la probabilidad de supervivencia en un grupo de sujetos en un intervalo de tiempo definido; y además, es el método más usado que tiene en cuenta las observaciones censuradas para estimar las funciones mencionadas anteriormente.

Si nuestro estudio no tuviese observaciones censuradas podemos estimar la función de supervivencia mediante la función de supervivencia empírica que dada una muestra de n observaciones obtenemos:

$$S_n(t) = \frac{1}{n} \quad \text{Card}\,\{i: t_i > t\}$$

Siendo:

- $Card(A)$ el cardinal o número de elementos de A.
- $Sn(t)$ es un estimador consistente de $S(t)$.

$$\forall t, \quad n\,Sn(t) \sim Bi[n, S(t)]; \qquad \forall t, \quad Sn(t) \xrightarrow[n\to\infty]{} N\left[S(t), \frac{S(t)(1-S(t))}{n}\right]$$

En el caso de que nuestro estudio tenga observaciones censuradas, el estimador de Kaplan-Meier es un poco más complejo.

Según la censura a la derecha, observamos dos pares formados por $(Y_i, \delta_i)\ i = 1 \ldots . n$

$$Y_i = \min\{T_i, C_i\} \quad \delta_i = \begin{cases} 1 & si \quad T_i \le C_i\ (observado) \\ 0 & si \quad T_i > C_i\ (censurado) \end{cases}$$

Los tiempos esperados observados permiten construir un conjunto $n' \le n$ intervalos.

$$I_i = \left(y_{(i-1)}, y_{(i)}\right) \qquad con\ i = 1, \ldots, n'$$

Sean los parámetros:

n_i = personas vivas y no censuradas justo antes de $y_{(i)}$.

d_i = personas que fallecen en $y_{(i)}$.

p_i $=Pr[sobrevivir\ a\ I_i \mid vivo\ al\ comienzo\ de\ I_i] = Pr[T > y_{(i)} \mid T > y_{(i-1)}]$.

q_i $= 1 - p_i = Pr[fallo\ en\ I_i \mid vivo\ al\ comienzo\ I_i] = Pr[T \le y_{(i)} \mid T > y_{(i-1)}]$.

A continuación, se recogen los estimadores de Kaplan-Meier (se notarán por estimadores K-M):

- ***Función de supervivencia***. Aplicando reiteradas veces la probabilidad condicional ($Pr[A \cap B] = Pr[A|B]Pr[B]$) se obtiene la siguiente expresión para $S(t)$.

$$S(t) = \Pr[T > t] = \prod_{i:\, y_{(i)} \le t} p_i$$

Por tanto, considerando los estimadores:

$$\widehat{q_i} = \frac{d_i}{n_i} \qquad \text{y} \qquad \widehat{p_i} = 1 - \widehat{q_i} = 1 - \frac{d_i}{n_i} = \frac{n_i - d_i}{n_i},$$

Se obtiene el estimador no paramétrico de la función de supervivencia propuesto por Kaplan-Meier (1958)

$$\hat{S}(t) = \prod_{i:\, y_{(i)} \le t} \widehat{p_i} = \prod_{i:\, y_{(i)} \le t} \frac{n_i - d_i}{n_i} = \prod_{i=1}^{k} \frac{n_i - d_i}{n_i}$$

$$k = y_{(k)} \le t < y_{(k+1)}$$

- ***La estimación de la tasa de fallo instantánea*** mediante el estimador de Kaplan-Meier en un intervalo, es el siguiente:

$$\hat{h}(t) = \frac{d_i}{n_i(t_{i+1} - t_i)} \qquad i: t_i \le t < t_{i+1}$$

- ***Estimación de fallo acumulada*** mediante los estimadores de Kaplan-Meier:

$$\widehat{H}(t) = -\ln\left[\hat{S}(t)\right] = -\ln \prod_{i:\, y_{(i)} \le t} \frac{n_i - d_i}{n_i}$$

$$\widehat{Var}\left[\widehat{H}(t)\right] = \sum_{i: y_{(i)} \le t} \frac{d_i}{n_i(n_i - d_i)}$$

- ***Estimación del tiempo media de vida truncada***:

Habitualmente, el tiempo medio de vida se estima a través de:

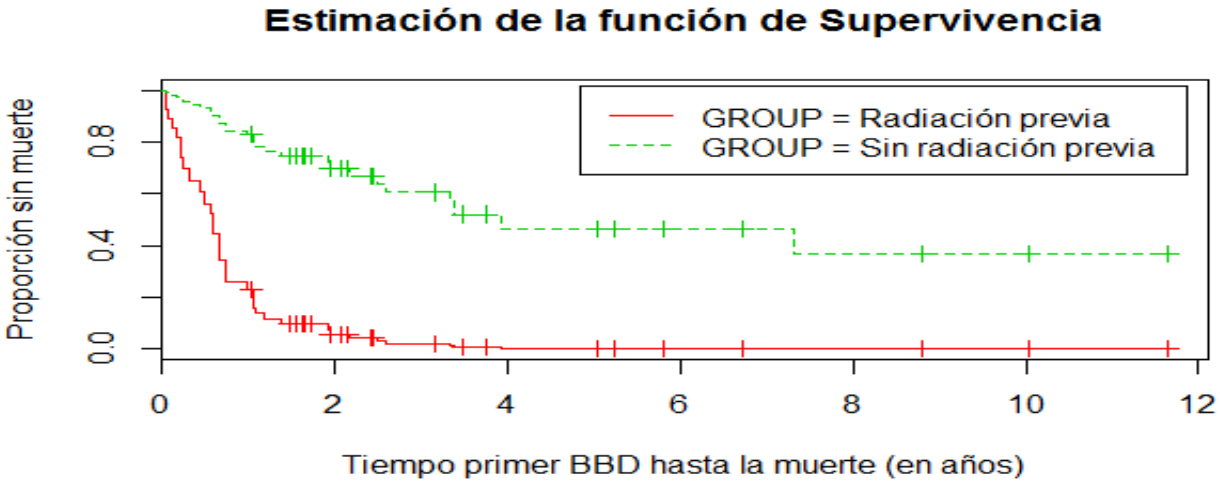
$$\hat{E}[T] = \int_0^{y_{(n)}} \hat{S}(t)dt$$

Pero en el caso de $y_{(n)} = max\{y_i\}$ no esté censurada, la integral anterior coincide con la que obtenemos en el intervalo $[0, \infty)$; pero en el caso de que esté censurada, el $\lim_{t \to \infty} \hat{S}(t) \neq 0$, por lo que la integral en el intervalo definido anteriormente no está definida. Para intentar evitar este problema, consideramos $y_{(n)}$ como si no

estuviera censurada obteniendo un estimador de la media, sesgado, el cual coincide con la curva de Kaplan-Meier:

$$\hat{E}[T] = \sum_{i=1}^{n'} [y_{(i)} - y_{(i-1)}]\hat{S}(y_i)$$

Otro problema de interés en el análisis de supervivencia que se aborda a través de métodos no paramétricos es la comparación de curvas de supervivencia. Se trata de contrastar la igualdad de dos funciones de supervivencia, o bien, la igualdad de dos funciones de distribución de tiempos de supervivencia F_1 y F_2, siendo la hipótesis nula $H_0: F_1 = F_2$. En la gráfica siguiente se ilustra una situación como la anteriormente descrita, con la representación de las estimaciones de Kaplan-Meier.



Para esta comparación, el método no paramétrico más utilizado es la construcción de tablas de doble entrada para cada fallo que observemos en nuestro estudio, siempre considerando ambos grupos.

Para ello se procede como sigue a continuación.

En primer lugar, suponemos que los tiempos observados son $t_{(1)} \leq t_{(2)} \leq \ldots \leq t_{(N)}$. Obteniendo en el tiempo $t_{(j)}$ el total de $d_j = d_{j1} + d_{j2}$ fallos en ambos grupos y expresando el número total de casos en riesgo de la siguiente forma $n_j = n_{j1} + n_{j2}$.

Tiempo: $t_{(j)}$	GRUPO 1	GRUPO 2	TOTAL
Fallos	d_{j1}	d_{j2}	d_j
No fallos	$n_{j1} - d_{j1}$	$n_{j2} - d_{j2}$	$n_j - d_j$
En riesgo	n_{j1}	n_{j2}	n_j

Bajo la hipótesis nula, una vez que tenemos fijadas las marginales, la variable X_j "$nº\ de\ fallos\ en\ el\ grupo\ 1\ en\ el\ instante\ j$" se distribuye según una distribución hipergeométrica:

$$X_j \big| n_j, n_{j1}, d_j \sim H\big[n_j, n_{j1}, d_j\big]$$

Por lo tanto, obtenemos las siguientes media y varianza:

$$E[X] = \frac{n_{j1} d_j}{n_j},$$

$$Var[X] = \frac{n_j - d_j}{n_j - 1} \frac{n_{j1} d_j}{n_j} \left(1 - \frac{n_{j1}}{n_j}\right) = \frac{n_{j1} n_{j2} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$$

Las distribuciones condicionadas son asintóticamente independientes

$$X_j \big| r\big(t_{(j)}\big), r_1\big(t_{(j)}\big), D_j \qquad j = 1 \dots N$$

Tanto los valores observados como los esperados podemos compararlo con el siguiente estadístico, convirtiéndolo en una distribución chi-cuadrado asintóticamente:

$$Z = \sum_{j=1}^{N} \left(d_{j1} - \frac{n_{j1} d_j}{n_j}\right) \Bigg/ \left(\sum_{j=1}^{N} \frac{n_{j1} n_{j2} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}\right)^{1/2}$$

O por una ponderación adecuada $w(t_{(j)})$

$$G = \sum_{j=1}^{N} w(t_{(j)}) \left(d_{j1} - \frac{n_{j1} d_j}{n_j}\right) \Bigg/ \left(\sum_{j=1}^{N} w^2(t_{(j)}) \frac{n_{j1} n_{j2} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}\right)^{1/2}$$

Bajo H_0: $G \sim \mathcal{X}_1^2$. Diversos estadísticos propuestos bajo este esquema de ponderación son los siguientes:

- $w(t_{(j)}) = 1$ para todo j=test de log-rangos de Mantel-Haenszel.
- $w(t_{(j)}) =$ n_j para todo j= test de Wilcoxon generalizado.
- $w(t_{(j)}) =$ raíz cuadrada (n_j) para todo j=test de Tarone y Ware.
- $w(t_{(j)}) = [\hat{S}(t_{j-1})]^p$ para todo j=test de Fleming-Harrington siendo $\hat{S}(t_{j-1})$ un estimador de K-M.
    - Si p=0 coincide con el test de log-rangos de Mantel-Haenszel.
    - Si p=1 se conoce como la modificación de Peto del test de Wilcoxon y es más sensible a diferencias en los valores iniciales de la curva de supervivencia.

Un desarrollo más detallado de este problema de comparación de dos grupos puede verse en Tableman y Kim (2003).

I.3. Métodos paramétricos.

Las distribuciones de probabilidad más frecuentes usadas en los métodos paramétricos son:

- Distribución exponencial, $T \sim Exp(\lambda)\ con\ \lambda > 0$:
    - $f(t) = \lambda e^{-\lambda t}$.
    - $S(t) = e^{-\lambda t}$.
    - $h(t) = \lambda$.
    - $H(t) = \lambda t$.
- Distribución gamma, $T \sim Ga(k, \lambda)\ con\ \lambda > 0$:
    - $f(t) = \frac{1}{\Gamma(k)} \lambda^k t^{k-1} e^{-\lambda t}$.
    - $S(t) y\ h(t)\ no\ tienen\ forma\ simple$.
- Distribución weibull, $T \sim W(\lambda, \alpha)\ con\ \lambda, \alpha > 0$.

- $f(t) = \lambda\alpha(\lambda t)^{\alpha-1}e^{-(\lambda t)^{\alpha}}$.
- $S(t) = e^{-(\lambda t)^{\alpha}}$.
- $h(t) = \lambda\alpha(\lambda t)^{\alpha-1}$.
- $H(t) = (\lambda t)^{\alpha-1}$.

- Distribución del valor extremo, tipo Gumbel, $EV(\mu, \alpha)\ con -\infty < \mu < \infty, \sigma > 0$.
  - $f(t) = \frac{1}{\sigma} exp\left[\frac{y-\mu}{\sigma} - \exp\left(\frac{y-\mu}{\sigma}\right)\right] \ -\infty < \mu < \infty$.
  - $S(t) = exp\left[-exp\left(\frac{y-\mu}{\sigma}\right)\right]$.
  - $h(t) = \frac{1}{\sigma}\exp\left(\frac{y-\mu}{\sigma}\right)$.
  - $H(t) = \exp\left(\frac{y-\mu}{\sigma}\right)$.
- Distribución log-normal, $T \sim LogN(\lambda, \alpha)\ con\ \lambda, \alpha > 0$.
  - $f(t) = (2\pi)^{-1/2}\alpha\, t^{-1}\, exp\left[-\frac{\alpha^2}{2} ln^2(\lambda t)\right]$ con $\Phi\, f.d.D.N.(0,1)$.
  - $S(t) = 1 - \Phi\,[\alpha \ln(\lambda t)]$.
- Distribución log-logística, $LogLog(\lambda, \alpha)\ con\ \lambda, \alpha > 0$.
  - $f(t) = \lambda\alpha(\lambda t)^{\alpha-1}[1 + (\lambda t)^{\alpha}]^{-2}$.
  - $S(t) = \frac{1}{1+(\lambda t)^{\alpha}}$.
  - $h(t) = \frac{\lambda\alpha(\lambda t)^{\alpha-1}}{1+(\lambda t)^{\alpha}}$.

La estimación de los parámetros de estas distribuciones se hace mediante máxima verosimilitud.

Los modelos de regresión al igual que en otros métodos estadísticos, pueden plantearse modelos de predicción del tiempo de fallo en función de otras variables potencialmente predictoras. En esta parte se presentan los principales modelos paramétricos de regresión en el análisis de supervivencia.

Todos estos modelos expresan la tasa de fallo instantánea en función de las variables predictoras.

- Modelo de regresión exponencial.

  Recuérdese que si $T \sim Exp(\lambda)$ con $\lambda > 0$ las funciones que se han definido previamente.

La tasa de fallo instantánea es constante respecto del tiempo. El modelo expresa esta tasa en función de una combinación lineal de las d variabels predictoras $\underline{X}$, usualmente la exponencial de dicha combinación, que asegura valores positivos:

$$h(t|x) = h_0(t)\exp(x^t\beta) = \lambda\exp(x^t\beta) = \lambda\exp(\beta_1 x_1 + \cdots + \beta_d x_d)$$

En general, la función $h_0(t)$ se denomina función de riesgo base, correspondiente a valores nulos de las covariantes.

$$\ln[h(t|x)] = \ln(\lambda) + \beta_1 x_1 + \cdots + \beta_d x_d$$

Es decir, las covariantes actúan de forma multiplicativa sobre la tasa de fallo instantánea y de forma aditiva sobre el logaritmo de dicha tasa.

$$S(t|x) = \{-h(t|x)t\} = \exp\{\lambda\, t\, e^{\beta_1 x_1 + \cdots + \beta_d x_d}\}$$

- Modelo de tiempo de fallo acelerado.

  Es un modelo de regresión log-lineal para T en el que se modeliza W como función lineal de las variables explicativas X.

$$W = \ln(T) = x^t\beta + Z^* \qquad Z^* \text{ con cierta distribución.}$$

  Así, $\exp\{x^t\beta\}$ actúa como un factor multiplicativo en el tiempo de supervivencia T.

$$T = \exp\{x^t\beta\}\exp\{Z^*\} = \exp\{x^t\beta\}T^* \text{ con } T^* = \exp\{Z^*\}$$

  El tiempo transformado $T^* = \exp\{Z^*\}$ dispone de una función de riesgo $h_0^*(t^*)$, que ha de ser no dependiente de las covariantes o variables explicativas.

- Modelo de riesgos proporcionales de Cox.

  Algunos de los modelos de regresión anteriores se pueden considerar casos particulares del modelo de riesgos proporcionales de Cox, en el que la función de fallo base $h_0(t)$ no está especificada.

$$h(t|x) = h_0(t)\exp(x^t\beta)$$

  En este modelo el cociente de riesgos (razón de riesgos, hazard ratios) no depende de t, conocida como propiedad de riesgos proporcionales.

$$\frac{h(t|x_1)}{h(t|x_2)} = \frac{\exp(x_1^t\beta)}{\exp(x_2^t\beta)} = \exp[(x_1 - x_2)^t\beta]$$

I.4. Aplicaciones ilustrativas generales.

Las técnicas y modelos incluidos en el tópico “Análisis de Supervivencia” son utilizados en muchas áreas: medicina, biología, industria,… Por tanto, es centro de interés científico tanto sus aplicaciones como el desarrollo y optimización de las técnicas ya propuestas.

Como ilustración puede considerarse diversos ejemplos de problemas de análisis de supervivencia (Kleinbaum & Klein, 2012).

- Estudio que sigue a pacientes con leucemia en remisión durante varias semanas para ver cuánto tiempo permanecen en remisión.
- Seguimiento a una cohorte libre de enfermedades de individuos durante varios años para ver quién desarrolla una enfermedad cardíaca.
- Seguimiento de individuos recién liberados bajo libertad condicional durante varias semanas para ver si son detenidos de nuevo. Este tipo de problema se denomina “estudio de la reincidencia”.
- Seguimiento de cómo los pacientes pueden sobrevivir mucho tiempo después de recibir un trasplante de corazón.

Todos los ejemplos anteriores son problemas de análisis de supervivencia debido a que la variable objetivo es el tiempo hasta que se produce un evento determinado. En el primer ejemplo, la participación de los pacientes con leucemia, el evento de interés (es decir, el fracaso) está "saliendo de la remisión", y el resultado es "tiempo en semanas hasta que una persona sale de la remisión." En el segundo ejemplo, el evento es "desarrollo enfermedades del corazón", y el resultado es "tiempo (en años) hasta que una persona desarrolle la enfermedad cardíaca." El tercero, un estudio sociológico más que médico, considera el caso de reincidencia y el resultado es "tiempo en semanas hasta una nueva detención." En el último ejemplo se ha considerado el evento "muerte", con el resultado "tiempo hasta la muerte (en meses después de haber recibido un trasplante)."

Siguiendo en esta línea de ilustrar aplicaciones del análisis de supervivencia, en el ámbito del marketing, se puede considerar el ejemplo de entrada de empresas en nuevos mercados (Fuentelsaz, Gomez, & Polo, 2004). En este estudio, su principal objetivo es dar respuesta a la siguiente pregunta:

“¿son diferentes los entrantes tempranos y los entrantes tardíos?”. Por lo tanto, el suceso de interés es la entrada de una caja de ahorros en una determinada provincia española.

Finalmente, también como ilustración de la amplitud de las aplicaciones de estas técnicas estadísticas, en el Capítulo III de la presente memoria se pueden ver otras ilustraciones con un mayor estudio de dicha aplicación.

Capítulo II. Modelo de Cox.

En este capítulo, se explica el tema principal de esta memoria, el modelo de Cox. En el apartado de estimación, se puede observar cómo se realiza la estimación mediante máxima verosimilitud, que en el modelo de Cox, se trata la verosimilitud "parcial" diferenciando si existen coincidencias en los tiempos de fallo o no; también se tratan los diferentes métodos que se pueden utilizar junto a la estimación de la función riesgo base y los métodos que se usan en R-Program. En el apartado de contraste de hipótesis, se explica los tres contrastes principalmente utilizados: contraste de Wald, razón de verosimilitudes y "score". Por último, se incluye el análisis del ajuste y diagnóstico del modelo, tratándose los procesos gráficos como son las curvas de supervivencia log-log y el contraste de bondad de ajuste.

II.1. Modelo

En el análisis de supervivencia el modelo de regresión más utilizado es el modelo Cox, dada su flexibilidad y dado que, a la hora de interpretar los coeficientes, es algo más simple que el resto de los modelos propuestos. Este modelo, también es denominado modelo de riesgos proporcionales.

Este modelo, trabaja primordialmente con la función de riesgo (hazard function) y es utilizado para detectar relaciones existentes entre el riesgo que se produce en un determinado individuo en el estudio y algunas variables independientes y/o explicativas; por lo que este modelo nos permite evaluar dentro de un conjunto de variables cuáles tienen relación, influencia…sobre la función de riesgo y por ello también en la función de supervivencia, ya que ambas funciones están conectadas.

Como se recoge en el aparado "I.3. Métodos paramétricos.", la función de riesgo en este modelo es la siguiente:

$$h(t|\underline{x}) = h_0(t)\exp(\underline{x}'\underline{\beta})$$

Tenemos un modelo semiparamétrico porque mientras que el riesgo basal, $h_0(t)$, puede tomar cualquier forma, las covariables entran a través de un modelo lineal con sus correspondientes parámetros. En dos puntos diferentes, $\underline{x_1}$ y $\underline{x_2}$, la proporción de riesgos para dichas observaciones son:

$$\frac{h(t|\underline{x_1})}{h(t|\underline{x_2})} = \frac{h_0(t)\exp(\underline{x_1}'\underline{\beta})}{h_0(t)\exp(\underline{x_2}'\underline{\beta})} = \frac{\exp(x_1^t\underline{\beta})}{\exp(x_2^t\underline{\beta})} = \exp[(\underline{x_1}' - \underline{x_2}')\underline{\beta}]$$

La expresión anterior es conocida como la hipótesis de riesgos proporcionales. Esta proporción también es conocida como la razón de riesgos. Podemos observar que es constante e independiente al tiempo t, por ello, el modelo también es conocido como modelo de riesgos proporcionales.

Por otro lado, la función de supervivencia viene dada por:

$$S(t|\underline{x}) = \exp\left(-\int_0^t h(u|\underline{x})du\right) = exp\left(-\exp\left(\underline{x}'\underline{\beta}\right)\int_0^t h_0(u)du\right)$$

$$= \left(\exp\left(-\int_0^t h_0(u)du\right)\right)^{\exp\left(\underline{x}'\underline{\beta}\right)} = (S_0(t))^{\exp\left(\underline{x}'\underline{\beta}\right)}$$

Mientras que la función de densidad de probabilidad de t dado x es:

$$f(t|\underline{x}) = h_0(t)\exp\left(\underline{x}'\underline{\beta}\right)(S_0(t))^{\exp\left(\underline{x}'\underline{\beta}\right)}$$

Dentro de este modelo encontramos dos generalizaciones importantes:

- $h_0(t)$ puede permitir que varíen los datos específicos en su subconjunto.
- $\underline{x}$ puede depender del tiempo, es decir, $\underline{x} = \underline{x}(t)$.

Ambas no son objeto de estudio en esta memoria.

II.2. Estimación.

A continuación describiremos como obtener las estimaciones de los parámetros del modelo de Cox, $\underline{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^t$ mediante máxima verosimilitud.

En el modelo de Cox, se utiliza la función de verosimilitud "parcial", porque la fórmula de probabilidad sólo considera probabilidades para aquellos sujetos que mueren/fallan y no considera las probabilidades para aquellos sujetos que son censurados. Sin embargo, en el cálculo de las probabilidades de los tiempos de muerte si tiene en cuenta todos los sujetos objeto de riesgo al inicio de los diferentes tiempos de muerte.

La verosimilitud parcial se puede expresar como el producto de diversas probabilidades, una para cada una de los *k* tiempos de fallo:

$$L = L_1 \times L_2 \times \dots \times L_k = \prod_{j=1}^{k} L_j$$

Por otro lado, para evitar la presencia de las funciones $h_0(.)$ y $H_0(.)$, el modelo de Cox propone una "verosimilitud parcial". Se parte de la idea de que los tiempos asociados a datos censurados ($\delta_i = 0$) no aportan información sobre el tiempo de fallo. En el caso de que no existan coincidencias de fallos, considera los tiempos de fallos observados ordenados:

$$t_1 < t_2 < \cdots < t_D$$

Con sus covariantes respectivamente asociadas,

$$x_{(1)}, x_{(2)}, \dots, x_{(D)}$$

Sea $R(t_{(i)})$ el conjunto de individuos en situación de riesgo en el instante del *i*-ésimo fallo (también incluye los casos censurados con censura mayor que t_i). Consideramos:

- Suceso "sobrevivir hasta el instante t_i" $\equiv V_i$
- Suceso "entre los casos $R(t_i)$ en riesgo hay un fallo en t_i" $\equiv M_i$
- Suceso "un individuo con $X = x_{(i)}$ falla en el instante t_i"$\equiv A_i$

Así, la verosimilitud parcial del *i*-ésimo individuo que falla es:

$$L^*_{(i)} = \Pr\left[\, A\left(x_{(i),}t_i\right)/\, V(t_i) \cap M(t_i)\right] = \Pr[A_i/V_i \cap M_i\,] = \frac{\Pr[\,A_i \cap V_i \cap M_i]}{\Pr[V_i \cap M_i]} =_{(1)} \frac{\Pr[A_i \cap V_i]}{A_i \cap M_i}$$

$$= \frac{\Pr[A_i/V_i]\,\Pr[V_i]}{\Pr[M_i/V_i]\,\Pr[V_i]} =_{(2)} \frac{\Pr[A_i/V_i]}{\Pr[M_i]} = \frac{h(t_i/x_{(i)},\beta)}{\sum_{j\epsilon R(t_i)} h(t_i/x_{(j)},\beta)}$$

$$= \frac{h_0(t_i)\exp\{x^t_{(i)}\beta\}}{\sum_{j\epsilon R(t_i)} h_0(t_i)\exp\{x^t_{(j)}\beta\}} = \frac{\exp\{x^t_{(i)}\beta\}}{\sum_{j\epsilon R(t_i)}\exp\{x^t_{(j)}\beta\}}$$

(1) Dado que $A_i \cap M_i = A_i$.

(2) Dado que M_i, V_i son independientes.

La expresión anterior puede ser considerada como una verosimilitud parcial que no depende de la función de riesgo base. La verosimilitud parcial conjunta de la muestra será:

$$L^*(\beta) = \prod_{i=1}^{D} L^*_{(i)} = \prod_{i=1}^{D} \frac{\exp\{x^t_{(i)}\beta\}}{\sum_{j\epsilon R(t_i)}\exp\{x^t_{(j)}\beta\}}$$

Y la log-verosimilitud parcial:

$$l^*(\beta) = \ln L^*(\beta) = \sum_{i=!}^{D}\left[x^t_{(i)}\beta - \ln\left(\sum_{j\epsilon R(t_i)}\exp\{x^t_{(j)}\beta\}\right)\right]$$

De este modo, los estimadores de máxima verosimilitud de β se obtienen maximizando la función de log-verosimilitud parcial, aplicando métodos iterados como el método de Newton-Raphson, que es llevado a cabo mediante derivadas parciales y resolviendo un sistema de ecuaciones:

$$\hat{\beta}: \qquad \frac{\partial\, l^*(\beta)}{\partial\, \beta} = 0$$

A partir de estos estimadores, se procede a la realización de los habituales test sobre los parámetros (test de razón de verosimilitudes, test de Wald y test de Score), basándose en la propiedad asintótica

$$\hat{\beta} \sim^a \mathcal{N}_p\left(\beta, \Phi^{-1}(\beta)\right),$$

siendo

$$\Phi(\beta) = -\frac{\partial^2 l^*(\beta)}{\partial \underline{\beta}\ \partial \underline{\beta}^t}$$

En el caso de que existan coincidencias en tiempos de fallo, usualmente se aplica uno de los tres siguientes métodos:

- Método de Breslow.
- Método de Efron.
- Método "Exact partial likelihood"

Sean:

- $U_{(i)}$ el conjunto de individuos que fallan en el instante t_i.
- $m_{(i)}$ la multiplicidad de t_i, es decir, el número de fallos en el instante t_i.

$$m_{(i)} = card\,\{U_{(i)}\}$$

- $r_i = card\,\{R(t_i)\}$.
- $z_{(i)} = \sum_{j\in U_{(i)}} x_{(j)}$.

Si el tiempo es "continuo", entonces la multiplicidades son pequeñas (generalmente iguales a 1, salvo en algunas excepciones). En tal caso se pueden aplicar los dos primeros métodos:

- Método de Breslow (1974), maximiza la verosimilitud parcial:

$$L_B^*(\beta) = \prod_{i=1}^{D} \frac{\exp\{z_{(i)}^t \beta\}}{\left[\sum_{j\epsilon R(t_i)} \exp\{x_{(i)}^t \beta\}\right]^{m_{(i)}}}$$

- Método de Efron (1977), maximiza la verosimilitud:

$$L_E^*(\beta) = \prod_{i=1}^{D} \frac{\exp\{z_{(i)}^t \beta\}}{\prod_{l=1}^{m_{(i)}} \left[\sum_{j\epsilon R(t_i)} \exp\{x_{(i)}^t \beta\} - \frac{j-1}{m_{(i)}} \sum_{j\epsilon U_{(i)}} \exp\{x_{(j)}^t \beta\}\right]}$$

Cuando los tiempos son observados en tiempo "discreto", los empates son "empates verdaderos", es decir, los fallos pasan realmente en el mismo instante de tiempo. Cox (1972) propone en este caso:

$$L_C^*(\beta) = \prod_{i=1}^{D} \frac{\exp\{z_{(i)}^t \beta\}}{\sum_{\boldsymbol{u}\in U_{(i)}} \exp\left\{\left(\sum_{j\in \boldsymbol{u}} x_{(j)}\right)^t \beta\right\}}$$

donde $U_{(i)}$ es el conjunto de todos los subconjuntos de $m_{(i)}$ sujetos que se pueden formar con los incluidos en el conjunto de riesgo $R(t_i)$.

En la estimación de la función de riesgo base, el modelo de regresión de Cox es considerado como un modelo semiparamétrico porque la función de riesgo base $h_0(t)$ se estima de forma "no paramétrica". En concreto se busca una función sólo con los valores no nulos en los tiempos de fallo observados, y así, $H_0(t)$ será una función escalonada.

Recuérdese que se observan

$$(y_j, \delta_j) \quad j = 1, \dots, n : \begin{cases} \text{Si } \delta_j = 1 \implies y_j \text{ tiempo de fallo } (t) \\ \text{Si } \delta_j = 0 \implies y_j \text{ tiempo de censura} \end{cases}$$

Se considera los tiempo de fallo (ordenados) $t_1 < t_2 < \cdots < t_D$, con covariantes asociadas, respectivamente, $x_{(1)}, x_{(2)}, \dots, x_{(D)}$.

La función de verosimilitud viene dada por:

$$L(y_1, \dots, y_n;\ h_0, \beta, x_1, \dots, x_n) = \prod_{j=1}^{n} f(y_j)^{\delta_j}\, S(y_j)^{1-\delta_j} = \prod_{j=1}^{n} h(y_j)^{\delta_j}\, S(y_j)$$

$$= \prod_{j=1}^{n} h(y_j)^{\delta_j} \exp\{-H(y_j)\} = \prod_{j=1}^{n} \left[h_0(y_j) \exp\{x_j^t \beta\}\right]^{\delta_j} \exp\{-H_0(y_j) \exp\{x_j^t \beta\}\}$$

$$= \left[\prod_{i=1}^{n} h_0(t_i) \exp\{x_{(i)}^t \beta\}\right] \left[\exp\left\{-\sum_{j=1}^{n} H_0(y_j) \exp\{x_j^t \beta\}\right\}\right]$$

Considerando $h_0(t_i) = h_{0_i} \quad i = 1, \dots, D$, $h_0(t) = 0\ \forall t$ no observado, con h_{0_i} desconocidas, que han de ser estimadas. La función acumulada será

$$H_0(y) = \sum_{i:\, t_i \le y} h_{0_i}$$

En consecuencia,

$$L(y_1, \dots, y_n;\ h_0, x_1, \dots, x_n) = \left[\prod_{i=1}^{D} h_{0_i} \exp\{x_{(i)}^t \beta\}\right]\left[\exp\left\{-\sum_{j=1}^{n} \exp\{x_j^t \beta\} \sum_{i:\, t_i \le y} h_{0_i}\right\}\right]$$

$$= \left[\prod_{i=1}^{D} h_{0_i} \exp\{x_{(i)}^t \beta\}\right]\left[\prod_{i=1}^{D} \exp\left\{-h_{0_i} \sum_{j:\, y_j \ge t_i} \exp\{x_j^t \beta\}\right\}\right]$$

Por otra parte, dado que

$$\{j: y_j \ge t_i\} = \{j: j \in R(t_i)\} = R(t_i)$$

$$L(y_1, \dots, y_n;\ h_0, x_1, \dots, x_n) = \prod_{i=1}^{D}\left[h_{0_i} \exp\{x_{(i)}^t \beta\} \exp\left\{-h_{0_i} \sum_{j \in R(t_i)} \exp\{x_j^t \beta\}\right\}\right]$$

Fijando β como el estimador de Máxima Verosimilitud Parcial obtenido anteriormente y tomando logaritmos:

$$l(h_0) = \sum_{i=1}^{D} \ln(h_{0_i}) + x_{(i)}^t \hat{\beta} - h_{0_i} \sum_{j \in R(t_i)} \exp\{x_{(i)}^t \hat{\beta}\}$$

Una buena estimación de los valores $h_{0_i}\ \ i = 1, \dots, D$ se puede obtener maximizando esta log-verosimilitud:

$$\frac{\partial}{\partial h_{0_i}} = \frac{1}{h_{0_i}} - \sum_{j \in R(t_i)} \exp\{x_{(i)}^t \hat{\beta}\} = 0 \Longrightarrow \hat{h}_{0_i} = \frac{1}{\sum_{j \in R(t_i)} \exp\{x_{(i)}^t \hat{\beta}\}}$$

$$\widehat{H_0}(y) = \sum_{i:\, t_i \le y} \hat{h}_{0_i} = \sum_{i:\, t_i \le y} \frac{1}{\sum_{j \in R(t_i)} \exp\{x_j^t \hat{\beta}\}}$$

En caso de empate, se usa el método de Nelson-Aalen-Breslow:

$$\widehat{H_0}(y) = \sum_{i:\, t_i \le y} \widehat{h_{0_i}} = \sum_{i:\, t_i \le y} \frac{m_{(i)}}{\sum_{j \in R(t_i)} \exp\{x_j^t \hat{\beta}\}}$$

Así, $\widehat{S_0}(y) = \exp\left[-\widehat{H_0}(y)\right]$ y por tanto $\hat{S}(y) = \left[\widehat{S_0}(y)\right]^{\exp(x^t \widehat{\beta})}$.

Los métodos de estimación de esta función no paramétrica utilizadas en el paquete *survival* de R son tres:

- Método Nelson-Aalen-Breslow.
- Método Efron:

$$\widehat{H_0}(y) = \sum_{i:\, t_i \le y} \frac{1}{\sum_{l \in R(t_i)} \exp\left\{x_l^t \hat{\beta} - \frac{j-1}{m_{(i)}} \sum_{l \in U_{(i)}} \exp\{x_{(l)}^t \beta\}\right\}}$$

La estimación también se puede complementar con el intervalo de confianza. Normalmente, el procedimiento que se utiliza para obtener un intervalo de confianza al 95% para la influencia de cada variable en la función de riesgo consiste en calcular la exponencial del intervalo de confianza asintótico obtenido para los parámetros o coeficientes del predictor lineal.

$$IC = \exp[\widehat{\beta_l} \pm 1.96\sqrt{\widehat{Var}\widehat{\beta_l}}]$$

Para mayor información, véase Kleinbaum & Klein (2012).

II.3. Contraste de hipótesis.

Como se ha recogido anteriormente, a partir de la función de verosimilitud parcial podemos obtener una estimación de los coeficientes $\underline{\hat{\beta}}$ cuya distribución es aproximadamente normal de media $\underline{\beta}$ y matriz de varianzas y covarianzas $\Sigma = \Phi^{-1}(\beta)$, que puede ser estimada por $\hat{\Sigma} = \Phi^{-1}(\underline{\hat{\beta}})$.

Para contrastar la hipótesis H_0: $\beta_j = 0 \ \ vs. \ H_1$: $\beta_j \neq 0$, es decir, la significación de la *j*-ésima covariante en el modelo, se puede utilizar el estadístico de Wald:

$$z = \frac{\widehat{\beta_J}}{\sqrt{\widehat{Var}\widehat{\beta_J}}}$$

Y para obtener el intervalo de confianza para el coeficiente $\widehat{\beta_J}$:

$$\widehat{\beta_J} \pm z_{\alpha/2}\sqrt{\widehat{Var}\widehat{\beta_J}}$$

Si se necesita hacer el test H_0: $\underline{\beta} = \underline{\beta_0}$ $vs. \ H_1$: $\underline{\beta} \neq \underline{\beta_0}$, se pueden usar tres contrastes:

1. Contraste de Wald, que se basa en $\underline{\hat{\beta}}$ sigue asintóticamente una distribución aproximadamente normal. Se considera el estadístico:

$$X_w = \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}\right)' \boldsymbol{\Phi}(\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})$$

que bajo hipótesis nula, sigue una distribución *chi*-cuadrado con p grados de libertad.

2. Contraste de la razón de verosimilitudes que compara el valor de la función de verosimilitud parcial evaluada en $\hat{\beta}, L^*\left(\hat{\beta}\right)$, con la verosimilitud parcial evaluada bajo hipótesis nula $L^*\left(\underline{\beta_0}\right)$;

$$X_{LR} = 2(\log L^*\left(\widehat{\boldsymbol{\beta}}\right) - \log L^*(\boldsymbol{\beta}_0))$$

Bajo hipótesis nula, el estadístico sigue una distribución *chi*-cuadrado con p grados de libertad.

3. Contraste "score" (Log Rank), que utiliza las derivadas del logaritmo de verosimilitud parcial evaluada en la hipótesis nula y supone que bajo hipótesis nula el vector scores:

$$X_s = \left(\frac{\partial L^*(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right)' \left(-\frac{\partial L^{*^2}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right)^{-1} \frac{\partial L^*(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}$$

El estadístico también sigue una distribución *chi*-cuadrado con p grados de libertad, bajo la hipótesis nula.

El contraste de Wald tiene una interpretación más directa que los otros dos contrastes, pero el contraste de Wald no es invariante ante parametrizaciones y los otros dos sí. Sin embargo, el test score para varios parámetros es más rápido computacionalmente y el test de razón de verosimilitudes converge más rápidamente a una distribución normal. La mayoría de las veces, se utiliza el contraste de razón de verosimilitudes.

II.4. Análisis del ajuste y diagnóstico del modelo.

Para evaluar la hipótesis básica del modelo de Cox, la hipótesis de riesgos proporcionales, existen diversos procedimientos. A continuación se recogen alguno de ellos.

1. Procedimiento gráfico.

Como procedimientos gráficos se dispone de dos tipos de gráficas, las cuales comparan las curvas de supervivencia log-log y las curvas de supervivencia esperadas contras las observadas. Recogemos a continuación las primeras.

- ***Curvas de supervivencia log-log***, se trata de una transformación de una curva de supervivencia estimada, que consiste en tomar el logaritmo natural en dos ocasiones. Matemáticamente, la curva log-log se determina por $-\ln(-\ln \hat{S})$, que puede ser tanto positiva como negativa.
  Por las propiedades del modelo,

$$\ln S(t, \boldsymbol{X}) = exp\left(\sum_{i=1}^{p} \beta_i X_i\right) \times \ln S_0(t)$$

  Dado que $0 \leq S(t, \boldsymbol{X}) \leq 1$, se tiene:

$$\begin{aligned} \ln[-\ln S(t, \boldsymbol{X})] &= ln\left[-exp\left(\sum_{i=1}^{p} \beta_i X_i\right) \times \ln S_0(t)\right] \\ &= ln\left[exp\left(\sum_{i=1}^{p} \beta_i X_i\right)\right] + \ln[-\ln S_0(t)] \\ &= \sum_{i=1}^{p} \beta_i X_i + ln[-\ln S_0(t)] \end{aligned}$$

  Esta expresión se puede expresar como la suma de dos términos, siendo uno el predictor lineal y el otro la transformación log-log de la función de supervivencia base.

Considerando dos especificaciones diferentes del vector X de variables predictoras, siendo estos $\mathbf{X}_1$ y $\mathbf{X}_2$, se tiene

$$\ln[-\ln S(t,\mathbf{X}_1)] = \sum_{i=1}^{p} \beta_i X_{1i} + \ l\ln[-\ln S_0(t)]$$

$$\ln[-\ln S(t,\mathbf{X}_2)] = \sum_{i=1}^{p} \beta_i X_{2i} + \ l\ln[-\ln S_0(t)]$$

Por tanto,

$$\ln[-\ln S(t,\mathbf{X}_1)] - (\ \ln[-\ln S(t,\mathbf{X}_2)]) = \sum_{i=1}^{p} \beta_i (X_{1i-}(X_{2i})$$

Como se puede observar, la función de supervivencia base se ha eliminado, por lo que la diferencia de las transformaciones log-log no depende de t.

Por otro lado, si se usa el álgebra, se puede escribir la ecuación anterior como la expresión de la curva de supervivencia log-log de la primera persona como la curva logarítmica doble para la segunda persona más el término de suma lineal independiente de t.

$$\ln[-\ln S(t,\mathbf{X}_1)] = \ln[-\ln S(t,\mathbf{X}_2)] + \sum_{i=1}^{p} \beta_i (X_{1i-}(X_{2i})$$

Si se representa gráficamente esta ecuación en el modelo de Cox, las curvas de supervivencia log-log estimadas para los individuos serían aproximadamente paralelas, es decir, en general, la distancia vertical entre las dos curvas debe ser aproximadamente constante.

Si un modelo de Cox es apropiado para un conjunto de predictores, las gráficas empíricas esperadas de las curvas de supervivencia log-log para diferentes individuos serán aproximadamente paralelas.

2. Contraste de bondad de ajuste.

Se han propuesto diferentes contrastes para estudiar la hipótesis de riesgos proporcionales. Uno de los más utilizados es el test propuesto por Harrell y Lee (1986).

Para cada predictor incluido en el modelo se consideran los determinados residuos de Schoenfeld (1982) para todos aquellos casos no censurados (es decir, con suceso o muerte). Para la *k*-ésima variable explicativa o predictor X_k se define:

$$r_{ik}^{(s)} = x_{ik} - \frac{\sum_{j \in R(t_i)} e^{\underline{x}_j^t \widehat{\underline{\beta}}} x_{jk}}{\sum_{j \in R(t_i)} e^{\underline{x}_j^t \widehat{\underline{\beta}}}} = x_{ik} - \hat{\bar{x}}_{(w_i)k}$$

Los residuos de Schoenfeld se pueden considerar como los valores observados menos los valores esperados de las covariantes en cada instante de fallo. Así, se tienen *p* series de residuos, uno para cada variable predictora $(k = 1,2, \dots, p)$. Si estos residuos mantienen un patrón aleatorio, es decir, no sistemático, proporciona una evidencia de que el efecto de la covariable no cambia respecto del tiempo, algo que presupone el modelo de Cox. Si hay algún tipo de patrón sistemático, sugiere que el efecto de la covariable cambia a lo largo del tiempo.

Así, si es cierta la propiedad de riesgos proporcionales, los residuos no mostrarán tendencias temporales y en el plot de los residuos frente al tiempo, la pendiente debe ser nula. Por tanto, estos plots

$$\left\{\left(t_i, r_{i,k}^{(s)}\right)\right\}_i$$

pueden utilizarse también como diagnósticos gráficos.

Por otra parte se puede realizar un contraste sobre la significación de la correlación entre los residuos y los rangos de los tiempos de fallo, según su media, en el siguiente proceso:

- Paso 1: Ejecutar el modelo PH de Cox y obtener los residuos de Schoenfeld para cada predictor. Para cada $k = 1, \dots, p$, $\left\{r_{i,k}^{(s)}\right\}_i$.
- Paso 2: Crear una variable que ocupa el orden de los tiempos de fallo. El sujeto que tiene el primer evento obtiene un valor de 1, el siguiente consigue un valor de 2 y así, sucesivamente.

  $\mathcal{U}_i =$ rango de t_i en la colección de tiempos de fallo $\{t_1, t_2, \dots\}$.

- Paso 3: Prueba de la correlación entre las variables creadas en la primera y segunda etapas. La hipótesis nula es que la correlación entre los residuos de Schoenfeld y tiempo de fallo censurado es cero.

$$H_0: \rho_{[k]} = \rho\left[r_k^{(s)}, \mathcal{U}\right] = 0$$

A partir de la estimación muestral de $\rho_{[k]}$ obtendrá con ella la correlación muestral asociada a los pares $\left\{\left(\mathcal{U}_i, r_{i,k}^{[s]}\right)\right\}_i$.

El rechazo de la hipótesis nula lleva a la conclusión de que se viola es supuesto de riesgos proporcionales.

Capítulo III. Aplicaciones.

En este último capítulo, por un lado se incluyen varias aplicaciones ilustrativas del modelo de Cox, recogiéndose, junto con el principal objetivo de su estudio, las variables estudiadas y una breve descripción de las mismas. En el segundo apartado, Modelo de Cox en R, se incluye un resumen de los procedimientos para abordar el análisis de este modelo en R-Program, ilustrándose con un ejemplo completo de todo lo estudiado en esta memoria.

III.1. Aplicaciones ilustrativas del Modelo de Cox.

Como se recoge en el primer capítulo, el modelo de Cox tiene múltiples aplicaciones, especialmente en el ámbito biosanitario. A continuación se recogen algunos trabajos publicados que nos permite ilustrar esta afirmación.

- **Modelo de regresión de Cox de la pérdida auditiva en trabajadores expuestos a ruido y fluidos de mecanizado o humos metálicos.** (Conte, Dominguez, Garcia Felipe, Rubio, & Perez Galdos, 2010)

  Este trabajo aborda el estudio del ruido, en la rama del metal, además de la presencia común de contaminantes químicos y físicos. Trata de analizar ambos junto con algunos hábitos personales para poder ver la influencia en la pérdida auditiva laboral.

  Para realizar dicho estudio, los autores recogen una muestra de 558 trabajadores y se aplica el modelo de regresión de Cox, principalmente con una finalidad explicativa. Define el carácter de las relaciones existentes entre las variables consideradas con respecto a tres situaciones causa-efecto:

  - Sano/Alterado.
  - Recuperable/No recuperable.
  - Sin caídas en conversacionales/Con caídas en conversacionales.

Las variables consideradas son:

- Intensidad de la exposición acústica en el puesto de trabajo.
- Presencia de agentes químicos.
- Tiempo total de exposición a ruido del trabajador.
- Estado audiométrico.
- Hábito de fumar.
- Exposición a ruido extralaboral.
- Uso de protección auditiva.

Los resultados del estudio reflejan que los fluidos de mecanizado retrasan la adquisición de los grados de alteración auditiva en presencia del ruido, mientras que los humos metálicos adelantan la adquisición. Por otro lado, la exposición al ruido extralaboral influye en la adquisición de un trauma acústico avanzado, el hábito de fumar es influyente en la adquisición de un trauma inicial acústico y por último, los equipos de protección auditiva son protectores del ruido pero no de la *ototoxicidad* de los humos metálicos.

- **Análisis de supervivencia aplicado al estudio de la mortalidad en injertos de *inchi*.** (García Bolívar, 2012)

  El principal objetivo de este estudio es la mortalidad de los injertos realizados en plantas de *inchi* (planta semileñosa y perenne, cuyos frutos son ricos en proteínas, aminoácidos, vitamina E y ácidos grasos esenciales, como omegas 3, 6, y 9) ya que al propagar el *inchi* pueden obtener mejoras en el rendimiento de los árboles madres.

  Para realizar el estudio, utilizaron plantas situadas en Maracay, estado Aragua (Venezuela), recolectando las semillas el mismo día de frutos en calderas, colocándose dichas semillas en propagadores de 2mx1m donde ya se encontraba el sustrato.

  El estudio constaba de 360 plantas, evaluándose el tiempo de mortalidad del injerto considerando después de 340 días con presencia de injerto muerto o injerto pegado.

  Las variables estudiadas son:

- Sexo de la planta donadora.
- Dosis de nitrógeno aplicada a patrones.
- Tiempo de remoción del plástico que cubre el injerto.

La variable objetivo o dependiente es el tiempo de supervivencia (en días).

El resultado obtenido del estudio en primer lugar, confirma que el sexo de la planta donadora no tiene ningún efecto sobre la mortalidad, en cambio con la remoción del plástico existe mayor riesgo de mortalidad a los 60 días.

- **Relación entre salud y renuncia al empleo en trabajadoras de la industria maquiladora electrónica de Tijuana.** (Guendelman, Samuels, & Ramirez-Zetina, 1999)

  El estudio consiste en el análisis de los factores de salud, laborales y sociales que contribuyen a renunciar al trabajo en dos maquiladores transnacionales del ramo electrónico de Tijuana (México). Las empresas maquiladoras son empresas que importan materiales sin pagar aranceles, su producto se comercializa en el país de origen de la materia prima y el proceso de fabricación se realiza en México.

  Para ello, el estudio consta de 725 mujeres empleadas en una planta estadounidense y otra japonesa.

  La muestra principal fue estratificada en dos intervalos ≤30 y >30 días laborales. El estudio fue seguido hasta la renuncia o el final del periodo de observación. La variable dependiente es la renuncia voluntaria o involuntaria al trabajo en la maquiladora.

  Las variables de estudio son:

  - Examen médico de ingreso.
  - Registro diario médico y de enfermería.
  - Reportes de accidentes y control de incapacidades.

  Por otro lado, en las mujeres que trabajaron más de 30 días también se analizaron otras variables:

  - Incapacidad por enfermedad general.

- Maternidad.
- Riesgos de trabajo.
- Número de días otorgados.

Dentro de las mujeres que renunciaban, fueron analizados su edad, estado de origen, estado civil, educación, características de salud, antecedentes, turno de trabajo, nacionalidad de la compañía…

Los datos fueron recogidos desde su primer día de trabajo hasta el día de la renuncia o final del periodo de observación.

Al final del estudio, un 17% renunció en su primer mes de empleo, un 54% en los 26 meses del estudio, en general, un 67% de renuncias en el primer año y un 81% en el segundo año.

Dentro de las renuncias, las variables que fueron determinantes para su despedida voluntaria o involuntaria fueron el turno de trabajo y la nacionalidad de la compañía. Comparando la nacionalidad de la compañía, vemos que la renuncia es mayor en la japonesa ya que es mucho más estricta y rigurosa.

Finalmente, a modo de ejemplo, se recoge otro estudio que no está en el ámbito biosanitario, para ilustrar la amplia gama de posibilidades de este modelo.

- **Análisis de la probabilidad condicional de incumplimiento de los mayores deudores privados del sistema financiero colombiano.** (Gómez González, Orozco Hinojosa, & Zamudio Gómez, 2006)

  El principal objetivo de este estudio es “encontrar los principales determinantes de la tasa de riesgo o probabilidad condicional de incumplimiento de las obligaciones financieras de las firmas del sector privado colombiano”.

  Este estudio se realiza a través de 2.000 deudores del sistema financiero colombiano.

  Las variables del estudio son:

  - Capitalización.
  - Calidad de los activos.

- Gerencia o eficiencia.
- Ganancias.
- Liquidez.

La conclusión del estudio es que el nivel de la deuda de las empresas es el principal determinante de la probabilidad condicional de incumplimiento. Además, es importante el efecto que tiene sobre esta probabilidad condicional pertenecer a algunos sectores económicos.

III.2. Modelo de Cox en R.

La función utilizada en el programa estadístico R para estudiar un modelo de regresión de Cox es *coxph()* que está implementada en el paquete de supervivencia "survival".

La función ajusta un modelo de regresión de riesgos proporcionales de Cox. Además incorpora otros aspectos tales como variables dependientes del tiempo, estratos dependientes del tiempo, múltiples eventos por cada sujeto, y otras extensiones.

El formato de la orden es:

```
coxph(formula, data=, weights, subset,
        na.action, init, control,
        ties=c("efron","breslow","exact"),
        singular.ok=TRUE, robust=FALSE,
        model=FALSE, x=FALSE, y=TRUE, tt, method, ...)
```

Los distintos argumentos de la formula son los siguientes:

- ***formula***: objeto "*formula*", con la respuesta a la izquierda de un operador ~, y las variables predictoras a la derecha. La respuesta debe ser un objeto supervivencia devuelto por la función "*Surv(var_tiempo, var_censura)*".
- ***data***: un data.frame en el que contenga las variables mencionadas en la fórmula.
- ***weights***: vector de ponderaciones de los casos.
- ***subset***: expresión que indica qué subconjunto de los datos se debe utilizar en el ajuste. Por defecto se incluyen todas las observaciones.
- ***na.action***: función de filtro de datos "missing"; por defecto no son considerados en el modelo. Esto se aplica a *model.frame* después de cualquier argumento subconjunto se ha utilizado. Por defecto es *option()$na.action*.
- ***init***: vector de los valores iniciales de la iteración. El valor inicial por defecto es cero para todas las variables.
- ***control***: objeto de la clase *coxph.control* especificando parámetros de control de la iteración. Más detalles en el manual (Therneau, 2016)
- ***ties***: una cadena de caracteres que especifica el método para el caso de empates. Si no hay empates en los tiempos de muerte todos los métodos son equivalentes. Aunque casi todos los programas de regresión de Cox utilizan el método de Breslow por defecto, no es así en este caso. Se utiliza como el valor por defecto la aproximación Efron (más eficiente computacionalmente). Es apropiado cuando los tiempos son un pequeño conjunto de valores discretos. Las opciones son: "efron", "breslow", "exact".
- ***singular.ok***: valor lógico que indica cómo manejar la colinealidad en la matriz de modelo. Si es "TRUE", el programa se saltará automáticamente las columnas de la matriz *X* que son combinaciones lineales del resto de las columnas. En este caso los coeficientes para tales columnas serán "NA", y la matriz de varianza contendrán ceros. Para los cálculos auxiliares, como el predictor lineal, los coeficientes que faltan son tratados como ceros.
- ***x***: valor lógico: si es "TRUE", la matriz *X* se incluye en el objeto resultado.
- ***y***: valor lógico: si es "TRUE", el vector de respuesta se incluye en el objeto resultado.
- Más detalles sobre otras opciones en el manuel (Therneau, 2016).

III.2.1. Ilustración.

Los datos provienen de un estudio clínico observaciones realizado en la Oregon Health Sciences University(OHSU). Se utilizan como ilustración por Tableman y Kim (2003) y se han obtenido de la dirección web http://www.math.ac.il/~yekutiel/survival/cns2.xls

Se trataron 58 pacientes, sin SIDA, con linfoma en el sistema nervioso central (SNC), desde enero de 1982 a marzo de 1992. Un grupo de pacientes (Grupo 1, 19 pacientes) recibieron radiación craneal antes del tratamiento de quimioterapia para la remisión de alteración de la barrera hematoencefálica (BBBD). Al resto (Grupo 0, 39 pacientes) se le administró sólo el tratamiento BBBD.

Se registraron un total de 16 variables sobre la respuesta radiográfica del tumor y la supervivencia:

- ***PTNUMBER***: número identificativo del paciente
- ***Group***: 1 = radiación previa; 0 = No radiación previa
- ***Sex***: 1=mujer, 0=hombre
- ***Age*** : Edad (en años) en el instante del primer BBBD
- ***Status***: estado 1=muerte ;0=no muerte
- ***DxtoB3***: Tiempo (años) desde el diagnóstico hasta el primer BBBD
- ***DxtoDeath***: Tiempo (años) desde el diagnóstico hasta la muerte
- ***B3toDeath***: Tiempo (años) del primer BBBD a muerte
- ***KPSPRE***: Puntuación de rendimiento de Karnofsky antes del primer BBBD, valor numérico 0-100
- ***LESSING***: Cantidad de Lesiones; simple = 0; múltiple = 1
- ***LESDEEP***: Profundidad de Lesiones; superficial=0;profunda=1
- ***LESSUP***: Tipo de lesiones; supra = 0; infra = 1; ambas = 2
- ***PROC***: Tipo de Procedimiento: resección subtotal = 1; biopsia = 2, otros = 3
- ***RAD4000***: Radiacion 4000; Si=1; No=0
- ***CHEMOPRIOR***: Si=1;No=0
- ***RESPONSE***: Respuesta del tumor a la quimio—completa=1; parcial=2.

La variable de respuesta principal es el tiempo hasta la muerte. Hay dos de estos tiempos:

- tiempo hasta la muerte desde el primer diagnóstico (DxtoDeath)
- tiempo hasta la muerte desde la primera BBBD (B3toDeath).

Dado que el interés está en determinar si la terapia de radiación tuvo algún efecto de prolongación del tiempo hasta la muerte, se considera como objetivo el tiempo hasta la muerte desde la primera BBBD.

Algunas de las preguntas de interés son:

- ¿Hay alguna diferencia en los tiempos de supervivencia entre los dos grupos (radiación previa, no radiación previa)?
- ¿Los subconjuntos de covariables disponibles ayudan a explicar este tiempo de supervivencia?

En primer lugar, se obtienen los datos en un archivo ".csv" por lo que tiene que leerse en R y a continuación se muestran las cinco primeras observaciones:

```
> cns<- read.csv(file = "cns2.csv", header=T, dec=",", sep=";")
> cns[1:5,]
  PT.NUMBER GROUP SEX AGE STATUS DXTOB3 DXTODEATH B3TODEATH KPS.PRE. LESSING LESDEEP
1         1     1   0  41      1   0.21      0.92      0.73       75       0       1
2         2     1   1  61      1   0.67      1.71      1.06       50       0       1
3         3     1   0  43      1   0.50      2.96      2.48       90       0       0
4         4     1   0  18      1   0.65      4.02      3.38      100       0       1
5         5     1   0  37      0   0.06      8.83      8.81       95       0       1
  LESSUP PROC RAD4000 CHEMOPRIOR RESPONSE
1      0    1       0          0        1
2      1    2       1          1        2
3      0    2       1          0        1
4      1    2       1          1        1
5      0    2       1          0        1
>
```

Como se puede observar el individuo 1 era un hombre con 41 años, que ha muerto, con radiación previa, con 0.21 años desde el diagnóstico hasta el primer BBBD ,0.92 años desde el diagnóstico hasta la muerte, 0.73 años desde el primer BBBD a la muerte, 75 de puntuación en el rendimiento de Karnofsky antes del primer BBBD, con ninguna lesión profunda de tipo infra, cuyo procedimiento llevado es resección subtotal, sin radiación 4000, sin radiación previa y con una respuesta del tumor a la quimio completa.

Ahora, se pasa a realizar el modelo de Cox:

b3t.cox<-coxph(Surv(B3TODEATH,STATUS) ~ GROUP + SEX + AGE + KPS.PRE.

+ LESSING + LESDEEP + LESSUP + PROC

+ RAD4000 + CHEMOPRIOR , data=cns)

Para analizar la salida de este modelo, se solicita un resumen del modelo a través de la orden "*summary*".

```
> summary(b3t.cox)
Call:
coxph(formula = Surv(B3TODEATH, STATUS) ~ GROUP + SEX + AGE +
    KPS.PRE. + LESSING + LESDEEP + LESSUP + PROC + RAD4000 +
    CHEMOPRIOR, data = cns)

               coef exp(coef) se(coef)      z Pr(>|z|)
GROUP       2.07905   7.99688  0.78380  2.653  0.00799 **
SEX        -1.62114   0.19767  0.53600 -3.025  0.00249 **
AGE         0.03499   1.03561  0.01619  2.161  0.03067 *
KPS.PRE.   -0.04406   0.95690  0.01531 -2.877  0.00401 **
LESSING     0.83337   2.30105  0.44099  1.890  0.05879 .
LESDEEP     0.15728   1.17032  0.45500  0.346  0.72959
LESSUP     -0.55130   0.57620  0.36573 -1.507  0.13171
PROC       -0.08613   0.91747  0.39117 -0.220  0.82573
RAD4000    -1.06219   0.34570  0.72810 -1.459  0.14461
CHEMOPRIOR  1.24727   3.48084  0.53214  2.344  0.01908 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

            exp(coef) exp(-coef) lower .95 upper .95
GROUP          7.9969     0.1250   1.72089   37.1611
SEX            0.1977     5.0589   0.06914    0.5652
AGE            1.0356     0.9656   1.00326    1.0690
KPS.PRE.       0.9569     1.0450   0.92861    0.9861
LESSING        2.3010     0.4346   0.96952    5.4613
LESDEEP        1.1703     0.8545   0.47975    2.8549
LESSUP         0.5762     1.7355   0.28136    1.1800
PROC           0.9175     1.0899   0.42622    1.9750
RAD4000        0.3457     2.8927   0.08297    1.4403
CHEMOPRIOR     3.4808     0.2873   1.22666    9.8774

Concordance= 0.778  (se = 0.054 )
Rsquare= 0.438   (max possible= 0.987 )
Likelihood ratio test= 33.39  on 10 df,   p=0.0002343
Wald test            = 28.45  on 10 df,   p=0.001531
Score (logrank) test = 35.86  on 10 df,   p=8.907e-05
```

Con la orden anterior se obtiene:

- Una tabla con coeficientes estimados, errores de estimación y significación de cada uno (z es el test de Wald, asintóticamente normal bajo la hipótesis de nulidad del coeficiente)

Por otro lado, exp(coef) permite una interpretación de los efectos multiplicativos de las variables explicativas sobre la función de riesgo $h(t)$,y se acompaña de un intervalo de confianza al 95%.

En la columna $\Pr[>|z|]$ se puede observar las variables que son significativas o no. En este caso las variables más significativas son "group,sex,kps.pre. y chemoprior" y las variables que no son significativas son "lesdeep, lessup, rad400 y proc" que podríamos eliminarlas del modelo y este no sufriría ningún cambio.

- El test de razón de verosimilitudes, test de Wald y test score para la significación del modelo son asintóticamente equivalentes, con sus p-valores significativos. Por tanto, se puede afirmar que el modelo permite explicar la variable tiempo de supervivencia considerada. La interpretación más detallada del efecto de cada variable se puede hacen en función de sus coeficientes.

Más información que se puede extraer del objeto creado es la siguiente:

- La estimación de los coeficientes del modelo.

```
> b3t.cox$coefficients
      GROUP         SEX         AGE    KPS.PRE.     LESSING     LESDEEP      LESSUP 
 2.07905119 -1.62114053  0.03498864 -0.04405523  0.83336535  0.15727779 -0.55129717 
       PROC     RAD4000  CHEMOPRIOR 
-0.08613014 -1.06219486  1.24727345 
```

- La estimación de la matriz de covarianzas de los estimadores:

```
> b3t.cox$var
              [,1]         [,2]          [,3]          [,4]          [,5]          [,6]
 [1,]  0.614334665  0.022028276  2.462180e-03  1.015816e-03  0.0843942628 -0.039930565
 [2,]  0.022028276  0.287295098 -3.272466e-03  2.502555e-03 -0.0585692271 -0.066556280
 [3,]  0.002462180 -0.003272466  2.620711e-04  6.278858e-05  0.0009387642  0.001403366
 [4,]  0.001015816  0.002502555  6.278858e-05  2.344214e-04 -0.0011290188  0.002063256
 [5,]  0.084394263 -0.058569227  9.387642e-04 -1.129019e-03  0.1944697587  0.017177276
 [6,] -0.039930565 -0.066556280  1.403366e-03  2.063256e-03  0.0171772756  0.207021703
 [7,] -0.061065139  0.047667402  6.745383e-04  2.483334e-03 -0.0804666359 -0.014622742
 [8,]  0.076069384  0.043511238 -1.319502e-03  8.557040e-04 -0.0147717080 -0.016278658
 [9,] -0.461314091 -0.063192539 -3.362972e-04  1.015858e-04  0.0046013362  0.054916015
[10,]  0.003147800 -0.145240411  2.342473e-03 -3.703894e-03  0.0212173697 -0.007031002
              [,7]         [,8]          [,9]         [,10]
 [1,] -0.0610651391  0.076069384 -0.4613140912  0.003147800
 [2,]  0.0476674019  0.043511238 -0.0631925393 -0.145240411
 [3,]  0.0006745383 -0.001319502 -0.0003362972  0.002342473
 [4,]  0.0024833340  0.000855704  0.0001015858 -0.003703894
 [5,] -0.0804666359 -0.014771708  0.0046013362  0.021217370
 [6,] -0.0146227421 -0.016278658  0.0549160147 -0.007031002
 [7,]  0.1337570742 -0.007181158  0.0099420257 -0.023003529
 [8,] -0.0071811582  0.153012035 -0.0636351255 -0.062411968
 [9,]  0.0099420257 -0.063635126  0.5301352243 -0.043975628
[10,] -0.0230035290 -0.062411968 -0.0439756281  0.283175901
```

- El valor de la log-verosimilitud del modelo y del modelo bajo H_0: $\beta = 0$.

```
> b3t.cox$loglik
[1] -125.5591 -108.8648
```

- El valor del test “Score”.

```
> b3t.cox$score
[1] 35.85703
```

- El valor del test de “Wald”.

```
> b3t.cox$wald.test
[1] 28.44515
```

- Los valores centrados ajustados del predictor lineal. Solo se muestran los cinco primeros:

```
> b3t.cox$linear.predictors[1:5]
[1]  1.7784506  1.5061149 -0.1180033 -0.5800175 -0.3909335
```

- Las medias de las variables predictoras.

```
> b3t.cox$means
    GROUP       SEX       AGE  KPS.PRE.   LESSING   LESDEEP   LESSUP
0.3275862 0.3448276 50.2758621 80.7758621 0.4310345 0.6379310 0.2758621
     PROC   RAD4000 CHEMOPRIOR
1.8275862 0.2758621  0.2586207
```

- Número de observaciones.

```
> b3t.cox$n
[1] 58
```

- Número de eventos(fallos, muertes…).

```
> b3t.cox$nevent
[1] 36
```

- Número de iteraciones realizadas en el procedimiento iterado de estimación.

```
> b3t.cox$iter
[1] 5
```

- El método utilizado.

```
> b3t.cox$method
[1] "efron"
```

- Los residuos martingala del modelo. Sólo se muestran los cinco primeros. (Tableman & Kim, 2003)

```
> b3t.cox$residuals[1:5]
         1          2          3          4          5
-0.6239569 -0.6947180  0.3022940  0.3602314 -1.1850416
```

Para la obtención de la función de supervivencia para valores medios de las predictoras se ejectua el siguiente comando:

```
> fsuperv.b3t = survfit(b3t.cox)
```

Esta vez se van a ir mostrando los valores correspondientes uno a uno:

- Tamaño muestral.

```
> fsuperv.b3t$n
[1] 58
```

- Tiempos observados.

```
> fsuperv.b3t$time
 [1]  0.04  0.06  0.13  0.17  0.21  0.23  0.31  0.44  0.50  0.56  0.58  0.60  0.65
[14]  0.73  0.98  1.04  1.06  1.08  1.17  1.38  1.48  1.56  1.63  1.65  1.73  1.92
[27]  1.94  1.96  2.08  2.15  2.17  2.42  2.44  2.48  2.58  3.17  3.33  3.38  3.48
[40]  3.75  3.92  5.04  5.23  5.81  6.73  7.31  8.81 10.04 11.65
```

- Casos en riesgos en cada tiempo.

```
> fsuperv.b3t$n.risk
 [1] 58 56 55 54 53 51 50 49 48 47 46 45 44 41 38 37 36 33 32 31 30 29 28 27 25 24 23
[28] 22 21 20 19 18 17 16 15 14 13 12 11 10  9  8  7  6  5  4  3  2  1
```

- Número de eventos en cada instante de tiempo.

```
> fsuperv.b3t$n.event
 [1] 2 1 1 1 2 1 1 1 1 1 1 1 3 3 1 0 3 1 1 1 0 0 0 0 0 1 1 0 0 0 1 0 0 1 1 0 1 1 0 0 1
[42] 0 0 0 0 1 0 0 0
```

- Número de censuras en cada instante de tiempo.

```
> fsuperv.b3t$n.event
 [1] 2 1 1 1 2 1 1 1 1 1 1 1 3 3 1 0 3 1 1 1 0 0 0 0 0 1 1 0 0 0 1 0 0 1 1 0 1 1 0 0 1
[42] 0 0 0 0 1 0 0 0
```

- Función de supervivencia en cada instante de tiempo para los casos con las covariables indicadas.

```
> fsuperv.b3t$surv
 [1] 0.9839785 0.9752583 0.9660237 0.9568327 0.9361081 0.9236351 0.9102273 0.8960161
 [9] 0.8816804 0.8673708 0.8523538 0.8374327 0.7919777 0.7433527 0.7248285 0.7248285
[17] 0.6674016 0.6471210 0.6239640 0.6003882 0.6003882 0.6003882 0.6003882 0.6003882
[25] 0.6003882 0.5659957 0.5330608 0.5330608 0.5330608 0.5330608 0.4966826 0.4966826
[33] 0.4966826 0.4560777 0.4158732 0.4158732 0.3740411 0.3189658 0.3189658 0.3189658
[41] 0.2590659 0.2590659 0.2590659 0.2590659 0.2590659 0.1734409 0.1734409 0.1734409
[49] 0.1734409
```

- Tipo de censura utilizado.

```
> fsuperv.b3t$type
[1] "right"
```

- Errores estándar de la estimación.

```
> fsuperv.b3t$std.err
 [1] 0.01233622 0.01611452 0.01978267 0.02317819 0.03044577 0.03444048 0.03855792
 [8] 0.04279104 0.04693443 0.05097479 0.05516952 0.05927788 0.07171023 0.08511989
[15] 0.09033713 0.09033713 0.10744408 0.11368399 0.12088240 0.12835693 0.12835693
[22] 0.12835693 0.12835693 0.12835693 0.12835693 0.12835693 0.14174479 0.15483315 0.15483315
[29] 0.15483315 0.15483315 0.17186178 0.17186178 0.17186178 0.19436997 0.21899650
[36] 0.21899650 0.24807016 0.29441777 0.29441777 0.29441777 0.36189221 0.36189221
[43] 0.36189221 0.36189221 0.36189221 0.55381506 0.55381506 0.55381506 0.55381506
```

- Nivel de confianza utilizado para el cálculo de los intervalos de confianza.

```
> fsuperv.b3t$conf.int
[1] 0.95
```

- Límite superior del intervalo de confianza.

```
> fsuperv.b3t$upper
 [1] 1.0000000 1.0000000 1.0000000 1.0000000 0.9936684 0.9881348 0.9816811 0.9744052
 [9] 0.9666335 0.9585055 0.9496865 0.9406052 0.9114919 0.8783125 0.8652276 0.8652276
[17] 0.8238418 0.8086368 0.7907784 0.7721288 0.7721288 0.7721288 0.7721288 0.7721288
[25] 0.7721288 0.7472510 0.7220561 0.7220561 0.7220561 0.7220561 0.6956134 0.6956134
[33] 0.6956134 0.6675547 0.6388092 0.6388092 0.6082428 0.5680061 0.5680061 0.5680061
[41] 0.5265668 0.5265668 0.5265668 0.5265668 0.5265668 0.5135229 0.5135229 0.5135229
[49] 0.5135229
```

- Límite inferior del intervalo de confianza.

```
> fsuperv.b3t$lower
 [1] 0.96047270 0.94493717 0.92928461 0.91433785 0.88188207 0.86334548 0.84397447
 [8] 0.82393322 0.80419348 0.78490112 0.76499662 0.74557702 0.68813414 0.62913048
[15] 0.60721176 0.60721176 0.54066800 0.51786616 0.49233897 0.46684696 0.46684696
[22] 0.46684696 0.46684696 0.46684696 0.46684696 0.42870619 0.39353424 0.39353424
[29] 0.39353424 0.39353424 0.35464181 0.35464181 0.35464181 0.31159522 0.27073900
[36] 0.27073900 0.23001787 0.17911636 0.17911636 0.17911636 0.12745800 0.12745800
[43] 0.12745800 0.12745800 0.12745800 0.05857918 0.05857918 0.05857918 0.05857918
```

Con los valores que se han obtenido se puede crear una tabla con los valores de supervivencia estimados más los intervalos de confianza.

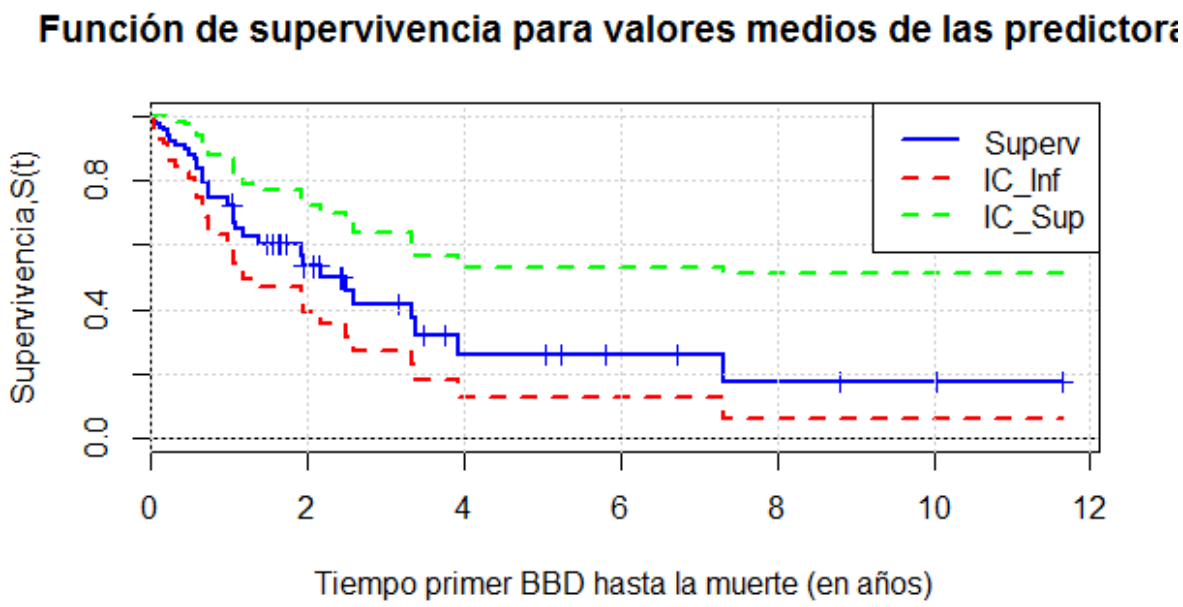
```
> t = cbind (fsuperv.b3t$lower,
+ fsuperv.b3t$surv, fsuperv.b3t$upper)
> t
           [,1]      [,2]      [,3]
 [1,] 0.96047270 0.9839785 1.0000000
 [2,] 0.94493717 0.9752583 1.0000000
 [3,] 0.92928461 0.9660237 1.0000000
 [4,] 0.91433785 0.9568327 1.0000000
 [5,] 0.88188207 0.9361081 0.9936684
 [6,] 0.86334548 0.9236351 0.9881348
 [7,] 0.84397447 0.9102273 0.9816811
 [8,] 0.82393322 0.8960161 0.9744052
 [9,] 0.80419348 0.8816804 0.9666335
[10,] 0.78490112 0.8673708 0.9585055
[11,] 0.76499662 0.8523538 0.9496865
[12,] 0.74557702 0.8374327 0.9406052
[13,] 0.68813414 0.7919777 0.9114919
[14,] 0.62913048 0.7433527 0.8783125
[15,] 0.60721176 0.7248285 0.8652276
[16,] 0.60721176 0.7248285 0.8652276
[17,] 0.54066800 0.6674016 0.8238418
[18,] 0.51786616 0.6471210 0.8086368
[19,] 0.49233897 0.6239640 0.7907784
[20,] 0.46684696 0.6003882 0.7721288
[21,] 0.46684696 0.6003882 0.7721288
[22,] 0.46684696 0.6003882 0.7721288
[23,] 0.46684696 0.6003882 0.7721288
[24,] 0.46684696 0.6003882 0.7721288
[25,] 0.46684696 0.6003882 0.7721288
[26,] 0.42870619 0.5659957 0.7472510
[27,] 0.39353424 0.5330608 0.7220561
[28,] 0.39353424 0.5330608 0.7220561
[29,] 0.39353424 0.5330608 0.7220561
[30,] 0.39353424 0.5330608 0.7220561
[31,] 0.35464181 0.4966826 0.6956134
[32,] 0.35464181 0.4966826 0.6956134
[33,] 0.35464181 0.4966826 0.6956134
[34,] 0.31159522 0.4560777 0.6675547
[35,] 0.27073900 0.4158732 0.6388092
[36,] 0.27073900 0.4158732 0.6388092
[37,] 0.23001787 0.3740411 0.6082428
[38,] 0.17911636 0.3189658 0.5680061
[39,] 0.17911636 0.3189658 0.5680061
[40,] 0.17911636 0.3189658 0.5680061
[41,] 0.12745800 0.2590659 0.5265668
[42,] 0.12745800 0.2590659 0.5265668
[43,] 0.12745800 0.2590659 0.5265668
[44,] 0.12745800 0.2590659 0.5265668
[45,] 0.12745800 0.2590659 0.5265668
[46,] 0.05857918 0.1734409 0.5135229
[47,] 0.05857918 0.1734409 0.5135229
[48,] 0.05857918 0.1734409 0.5135229
[49,] 0.05857918 0.1734409 0.5135229
```

Una vez que se ha obtenido la tabla anterior, se crea una gráfica con la función de supervivencia de los valores medios de las predictoras junto a sus intervalos de confianza.

- Plot “Función de supervivencia para los valores medios de las predictoras”

```
plot(survfit(b3t.cox),xlab="Tiempo primer BBD hasta la muerte (en años)",
     ylab="Supervivencia,S(t)", col=c("blue", "red", "green"),lwd=2,
main="Función de supervivencia para valores medios de las predictoras")
abline(h=0)
grid()
legend("topright",lty=c(1,2,2),col=c("blue", "red", "green"),lwd=2,
       legend=c("Superv","IC_Inf","IC_Sup"))
```

Obteniéndose la siguiente gráfica:



Para seguir profundizando en la interpretación del modelo ajustado se va a crear dos individuos nuevos idénticos exceptuando la variable “group”, es decir, un individuo habrá tenido radiación previa y otro no; con el resto de variables exactamente iguales.

Estos individuos son creados con los siguientes comandos:

```
cns.fin <- with(cns, data.frame(PT.NUMBER=c(1001,1002),
                                GROUP=c(1,0),
                                SEX=rep(0,2),
                                AGE=rep(b3t.cox$means[3],2),
                                KPS.PRE.=rep(b3t.cox$means[4],2),
                                LESSING=rep(1,2),
                                LESDEEP=rep(0,2),
                                LESSUP=rep(2,2),
                                PROC=rep(2,2),
                                RAD4000=rep(1,2),
                                CHEMOPRIOR=rep(1,2)))
```

Siendo los individuos creados:

```
  PT.NUMBER GROUP SEX      AGE KPS.PRE. LESSING LESDEEP LESSUP PROC RAD4000 CHEMOPRIOR
1      1001     1   0 50.27586 80.77586       1       0      2    2       1          1
2      1002     0   0 50.27586 80.77586       1       0      2    2       1          1
```

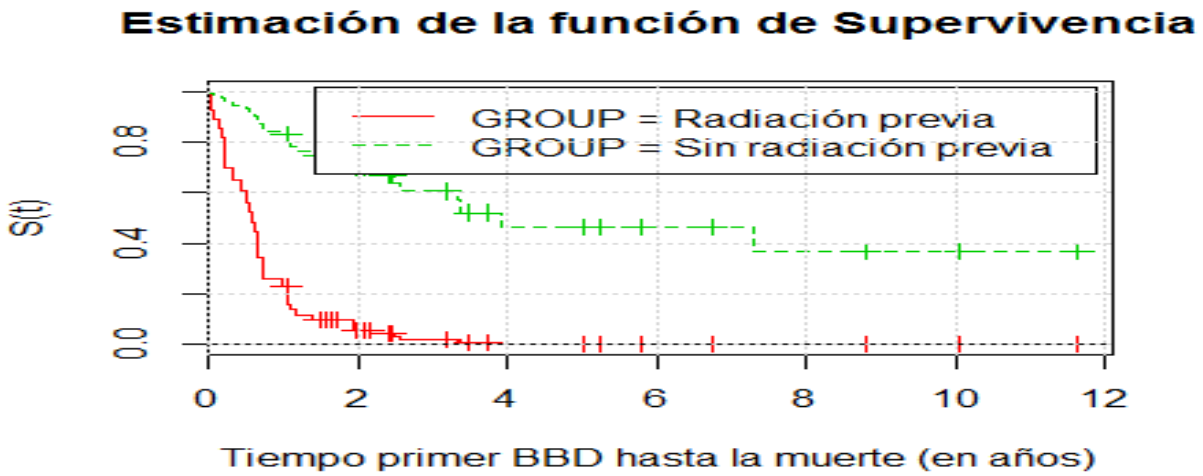
Es decir, ambos individuos son hombres, con una edad media de 50.276 años, con una puntuación media de 80.776 en el rendimiento de Karnofsky antes del primer BBBD, con lesiones múltiples, superficiales y de ambos tipo, supra e infra. Además, estos individuos tienen biopsia en el tipo del procedimiento, con radiación 4000 y con quimioterapia previa.

Para poder observar el estudio de estos nuevos individuos, se crea:

- Plot “Función de supervivencia”, ejecutado con los siguientes comandos.

```
plot(survfit(b3t.cox, newdata=cns.fin), conf.int=FALSE,
     lty=c(1, 2), ylim=c(0.0, 1), xlab="Tiempo primer BBD hasta la muerte (en años)",
     ylab="S(t)",col = c(2,3),
     main="Estimación de la función de Supervivencia")
legend("topright", legend=c("GROUP = Radiación previa", "GROUP = Sin radiación previa"),
       lty=c(1 ,2), col = c(2,3),inset=0.02)
abline(h=0)
grid()
```

Obteniéndose:

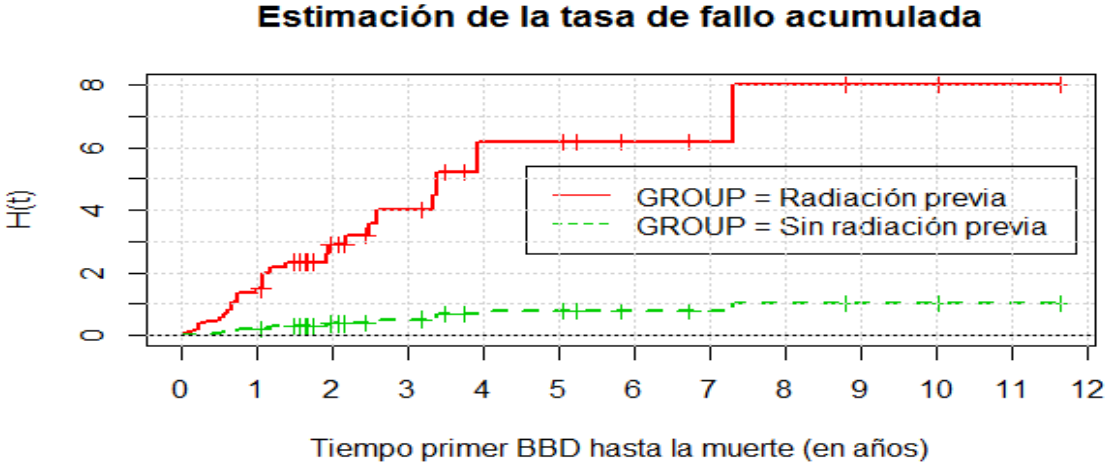


En esta gráfica, se puede observar que el individuo que no ha tenido radiación previa a su estudio, tiene más tiempo de vida, es decir, mayor probabilidad de supervivencia. Si nos centramos en el intervalo de 8-10 años se ve que mientras el individuo con radiación previa "ha fallecido", el individuo sin radiación previa tiene una probabilidad de supervivencia de 0.4.

- Plot "Función de la tasa de fallo acumulada".

  Para la realización de este plot, se usa los siguientes comandos:

```
plot(survfit(b3t.cox, newdata=cns.fin),conf.int=FALSE,lty=c(1, 2),
     xlab="Tiempo primer BBD hasta la muerte (en años)",
     ylab="H(t)",lab=c(10, 10, 7),lwd=2,fun="cumhaz", col = c(2,3),
     main="Estimación de la tasa de fallo acumulada")
legend("right", legend=c("GROUP = Radiación previa", "GROUP = Sin radiación previa"),
       lty=c(1 ,2), col = c(2,3),inset=0.02)
abline(h=0)
grid()
```



En esta gráfica, se puede observar lo mismo pero “invertido”, es decir, el individuo sin radiación previa tiene una tasa de fallo muy baja en comparación con el individuo que ha sufrido radiación previa. Si nos volvemos a centrar en el intervalo de 8-10 años, podemos observar que la tasa de fallo acumulada del individuo con radiación previa es de 8 mientras que la del individuo que no sufrido radiación previa es de 1.

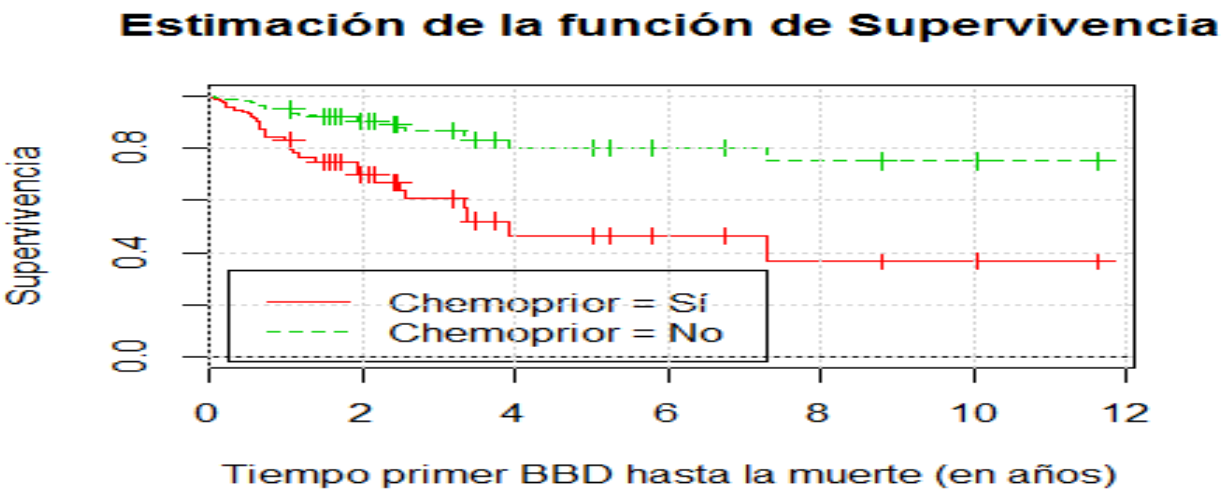
En estos individuos, se ha estudiado la variable “GROUP” que se puede observar anteriormente en el “summary” del modelo era significativa de dicho modelo, ahora se va a realizar el estudio con dos nuevos individuos pero comparando la variable “CHEMOPRIOR” que también es significativa del modelo.

Los nuevos individuos que se crean son:

	PT.NUMBER	GROUP	SEX	AGE	KPS.PRE.	LESSING	LESDEEP	LESSUP	PROC	RAD4000	CHEMOPRIOR
1	1003	0	0	50.27586	80.77586	1	0	2	2	1	1
2	1004	0	0	50.27586	80.77586	1	0	2	2	1	0

Y en este caso, hemos creado exactamente las mismas gráficas con los nuevos individuos.

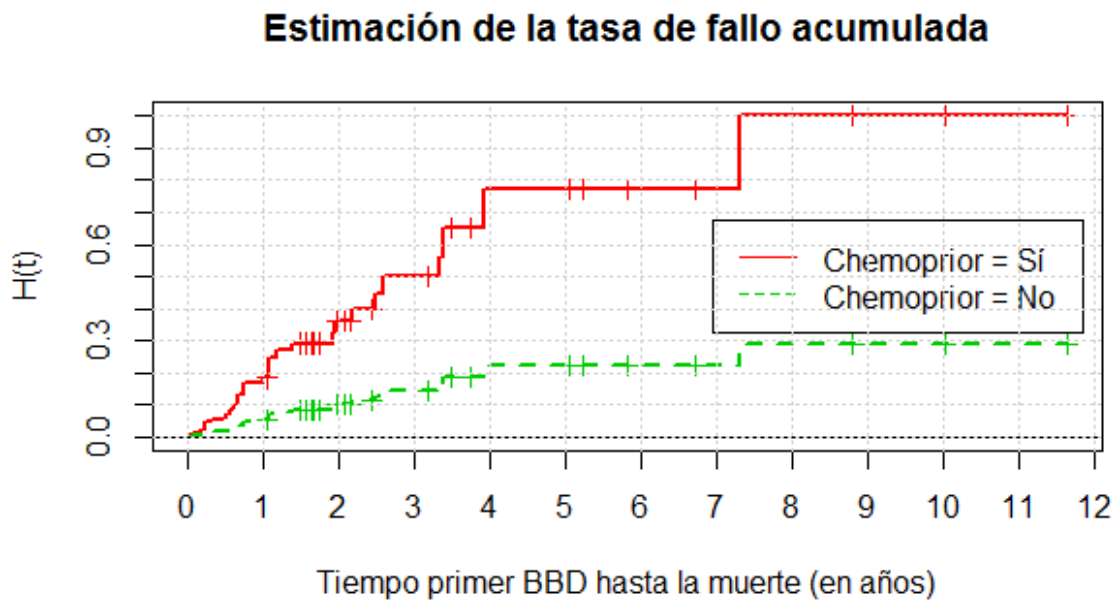
- Plot “Función de Supervivencia”



En esta gráfica, se puede observar que ambos individuos tienen una función de supervivencia similar; y con esta variable vuelve a pasar lo mismo que con la variable “group”; el individuo que ha sufrido quimioterapia previa tiene menor probabilidad de supervivencia. En el intervalo de 8-10 años, se puede observar que el individuo con quimioterapia previa tiene una probabilidad de supervivncia

aproximadamente de 0.4 mientras que el individuo que no ha sufrido quimioterapia previa tiene una probabilidad de supervivencia de 0.8.

- Plot “Función estimación de la tasa de fallo acumulada”



En la gráfica sobre la función de riesgo acumulado o tasas de fallo, vuelve a ocurrir lo mismo. Si nos centramos en el intervalo de 8-10 años, el individuo que ha sufrido quimioterapia previa tiene una tasa de fallo de 1 mientras que el individuo que no ha sufrido quimioterapia previa tiene un 0.3 aproximadamente.

Al principio de nuestra ilustración, se propusieron las siguientes preguntas.

- ¿Hay alguna diferencia en los tiempos de supervivencia entre los dos grupos (radiación previa, no radiación previa)?

  Al finalizar el estudiar y predecir dos nuevas variables se puede observar que los individuos que han sufrido radiación previa tienen una probabilidad de supervivencia menor a los individuos que no han sufrido dicha radiación, por lo que se podría decir que sí existe diferencia entre unos individuos y otros.

- ¿Los subconjuntos de covariables disponibles ayudan a explicar este tiempo de supervivencia?

  En general el modelo resulta significativo, por tanto las variables en su conjunto permiten explicar el comportamiento de la variable tiempo de supervivencia. En concreto, las variables que resultan significativas son radiación previa (sí/no), sexo, edad, rendimiento de Karnofsky antes del primer BBBD y quimioterapia previa (sí/no). Obviamente, la interpretación completa de la influencia de estas variables, dependiendo del signo de los coeficientes, debe ser realizada por un especialista en el área de investigación del tema tratado.

Referencias

Breslow, N. (1974). Covariance Analysis of Survival Data under the Proportional Hazards Model. *International Statistical Review*, 43-54.

Conte, J., Dominguez, A., Garcia Felipe, A., Rubio, E., & Perez Galdos, A. (2010). Modelo de regresión de Cox de la pérdida auditiva en trabajadores expuestos a ruidos y fluidos de mecanizado o humos metálicos. *An. Sist. Sanit. Navar. 33(1)*, 11-20.

Cox, D. (1972). Regression Models and Life Tables. *Journal of the Royal Statistical Society*, 187-220.

Efron, B. (1977). The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association*, 557-565.

Fuentelsaz, L., Gomez, J., & Polo, Y. (2004). Aplicaciones del análisis de supervivencia a la investigación en economía de la empresa. *Cuadernos de Economía y Dirección de la Empresa, 19*, 081-114.

García Bolívar, J. (2012). Análisis de supervivencia aplicado al estudio de la mortalidad en injertos de inchi.(Caryodendron orinocense Karsten). *Revista Científica UDO Agrícola 12(4)*, 759-769.

Gómez González, J., Orozco Hinojosa, I., & Zamudio Gómez, N. (2006). Análisis de la probabilidad condicional de incumplimiento de los mayores deudores privados del sistema financiero colombiano. In *Temas de Estabilidad Financiero* (pp. 93-102). Banco de la República (Colombia).

Guendelman, S., Samuels, S., & Ramirez-Zetina, M. (1999). Relación entre salud y renuncia al empleo en trabajadores de la industria maquiladora electrónica de Tijuana. . *Salud Pública México, Vol.41 (4)*, 286-296.

Harrell, F., & Lee, K. (1986). Verifying assumptions of the Cox proportional hazards model. *Proceedings of the Eleventh Annual SAWS User's Group International Conference* (pp. 823-828). Cary, NC: SAS Institute, Inc.

Kaplan, E., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of American Statistical Association*, 53: 457-481.

Kleinbaum, D., & Klein, M. (2012). *Survival Analysis: a Self-Learning Text. Third Edition.* Springer Science+Business.

Schoenfeld, D. (1982). Partial residuals for th proportionl hazards model. *Biometrika*, 51-55.

Tableman, M., & Kim, J. (2003). *Survival Analysis using S: Analysis of Time-of-Event data.* Chapman & Hall/CRC.

Therneau, T. (2016). *A package for Survival in S. R package version 2.39-4*. Retrieved from http://CRAN.R-project.org/package=survival