UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE CIENCIAS MATEMÁTICAS

UNIDAD DE POSTGRADO

Forma funcional de covariables en el modelo de Cox

TESIS

para optar el grado académico de Magíster en Estadística Matemática

AUTOR

Daniel Octavio Roque Roque

Lima-Perú

2009

“FORMA FUNCIONAL DE COVARIABLES EN EL MODELO DE COX”

DANIEL OCTAVIO ROQUE ROQUE

Tesis presentada a consideración del cuerpo docente de la Facultad de Ciencias Matemáticas, de la Universidad Nacional Mayor de San Marcos, como parte de los requisitos para obtener El Grado Académico de Magister en Estadística Matemática

Aprobada por:

Mg. Violeta Nolberto Sifuentes
Presidenta

Mg. Antonio Bravo Quiroz
Miembro-Asesor

Mg. Ysabel Adriazola Cruz
Miembro

Mg.Ysela Aguero Palacios
Miembro

Mg. Estela Ponce Aruneri
Miembro

LIMA - PERU

Noviembre-2009

DEDICATORIA

En memória de mi querido padre:
HIPOLITO
y, a mis seres queridos.

AGRADECIMIENTO

Al Magister Antonio Bravo Quiroz, no sólo por su impecable labor de dirección como asesor, sino también por sus continuas enseñanzas y apoyo en la elaboración del presente trabajo de Tesis..

A la Doctora Blanca Rosa Maquera Sosa, ex compañera de Pre-grado en la Universidad Nacional de San Agustín de Arequipa. Actual docente de la Universidad de Passo Fundo de la Ciudad de Passo Fundo-Brasil; por su respaldo personal y apoyo bibliográfico.

CONTENIDO

INTRODUCCIÓN

El análisis de supervivencia consiste en una colección de procedimientos estadísticos que permiten analizar y modelar, a los datos relacionados a la variable respuesta T, que a partir de un tiempo inicial pre-establecido; T representa el tiempo de seguimiento hasta la ocurrencia de un determinado suceso o evento de interés previamente fijado por el investigador, de modo que este evento de interés puede ser: muerte, fallo de un injerto renal, efectividad de un tratamiento, aparición de una complicación clínica, etc. Dichos procedimientos son hoy en día, una metodología fundamental en gran parte de los ensayos clínicos y de los estudios epidemiológicos que son experimentos de tipo longitudinal y prospectivo.

El análisis de supervivencia se aplica a los datos biomédicos obtenidos según un protocolo que consiste en definir de manera precisa el momento inicial de la observación y el momento final, ya que la variable aleatoria T representa el tiempo transcurrido entre el inicio del tratamiento u observación y la consecución de un cierto evento de interés llamado falla o muerte. Sin embargo, puede haber individuos que no presentan el evento respectivo pre-establecido mientras dure el periodo de seguimiento, a los cuales, se les denomina individuos censurados o datos censurados; por eso, el objetivo principal del análisis de supervivencia es incorporar a su análisis ésta información parcial proporcionada por los individuos censurados mediante métodos desarrollados para ese fin.

El origen del nombre de análisis de supervivencia se remonta al siglo XVII con la construcción de las tablas de vida dentro de las ciencias actuariales y, se debe fundamentalmente a que en muchas aplicaciones el evento de interés era la muerte.

Tal y como se ha descrito en los párrafos anteriores, en el análisis de supervivencia las variables de interés son los datos censurados “C” y el tiempo “T” hasta que ocurra un evento o suceso de interés. La ocurrencia de la variable respuesta T puede ser afectada por diferentes variables llamadas covariables o variables pronósticos que son propias de cada sujeto o individuo; así pues, se modela la función de riesgo como una función del tiempo y de las variables pronósticos, de modo que se obtiene lo que se llama el modelo de regresión de riesgos proporcionales; más conocido como el modelo de regresión de Cox (1972); que tiene una forma funcional log-lineal.

Un aspecto importante del modelo de Cox es que, nos permite predecir la función de supervivencia de cada sujeto utilizando los valores pronósticos o explicativas y la función de riesgo acumulado; para lo cual, las variables explicativas (covariables) para dos individuos diferentes se debe satisfacer la suposición de riesgos proporcionales que es un principio fundamental del modelo de Cox. Sin embargo, existen covariables que no cumplen con la suposición de riesgo proporcional, lo cual implica estratificar los datos de modo que la suposición de riesgos proporcionales sea válida en cada estrato. Por otro lado, pueden presentarse también casos en que los covariables cambian de valores durante el periodo de estudio; es decir, covariables tiempo dependientes, los cuales; tampoco admiten el cumplimiento de la suposición de riesgos proporcionales; obteniéndose así un modelo cuyas covariables son tiempo-dependientes.

Tanto la estratificación de los datos y así como los modelos tiempo-dependientes en el modelo de Cox son denominados modelos de forma funcional de las covariables, porque dichas variables explicativas tienen como resultados valores numéricas en función del tiempo o grupo.

Los objetivos de mi trabajo de Tesis son siguientes:

1. Presentar la modelación de la función de riesgo de cada individuo en función de la variable respuesta T y las variables explicativas (covariables).

2. Estudiar la forma funcional del modelo de Cox tanto por covariables dependientes del tiempo así como la estratificación de los datos.
3. Aplicar la forma funcional de las covariables del modelo de Cox al análisis de los datos de pacientes HIV y leucemia.

Para lograr los objetivos propuesto en el presente trabajo de tesis, la estructuración de los contenidos capitulares es como sigue:

El primer capítulo es de carácter introductorio al tema de análisis de supervivencia, se desarrolla los conceptos básicos, tales como: función de supervivencia, función de riesgo y tipos de censuramiento.

En el segundo capítulo, se desarrolla estimaciones desde el punto de vista no paramétrico; tales como: el estimador de Kaplan-Meier, estimación de la función de riesgo, varianza de la función de supervivencia y comparación de dos funciones de supervivencia.

En el capítulo III se desarrolla el modelo Cox, estimación de los parámetros del modelo de Cox, estimación de función de riesgo base, interpretación de los parámetros del modelo de Cox y la verificación de la suposición de los riesgos proporcionales mediante el análisis de los residuales.

En el capítulo IV se estudia la forma funcional de los covariables, tales como: modelo de Cox con covariables dependientes del tiempo, modelo de Cox estratificado y aplicaciones a pacientes con HIV y leucemia.

Debido a su gran preponderancia en diversas investigaciones científicas, y de fácil accesibilidad por el usuario se ha escogido el lenguaje R versión 2.4.0, como apoyo para realizar los cálculos y gráficos estadísticos en todo el trabajo de tesis.

CAPITULO I:

CONCEPTOS BASICOS DE ANALISIS DE SUPERVIVENCIA.

1.1.- INTRODUCCIÓN.

La presencia de información incompleta o parcial representada por datos censurados, hace difícil su análisis y manejo de los datos de supervivencia mediante los métodos de la estadística clásica; ya que, ignorarlos estos datos como datos faltantes, sería desconocerlos el aporte parcial de estos individuos al estudio. Por otro lado, el hecho de desconocerlos los individuos censurados estaríamos propiciando una metodología contrario a la filosofía de la estadística de incorporar toda la información disponible dentro del análisis.

1.2.- FUNCION DE SUPERVIVENCIA Y RIESGO.

En análisis de supervivencia es habitual utilizar otras funciones, además de las funciones de distribuciones o densidad, para caracterizar la distribución de probabilidad de la variable aleatoria T, que representa el tiempo desde el inicio de la observación de un experimento hasta la ocurrencia de un evento o suceso de interés, en dicha observación.

El comportamiento de la variable aleatoria $T \geq 0$ (tiempo de supervivencia), puede ser expresado a través de varias funciones matemáticamente equivalentes, tales que, si una de ellas es especificada, las otras pueden ser derivadas. Entre ellas tenemos, la función de densidad de probabilidad f(t), la función de supervivencia denotado por S(t) y, la función de riesgo h(t); que

serán descritas en detalle en este capítulo. Estas tres funciones son utilizadas en la práctica para describir los diferentes aspectos que presentan el conjunto de datos en el análisis de supervivencia.

DEFINICIÓN 1.2.1.- Un tiempo de falla (tiempo de supervivencia, o tiempo de vida) T, es una variable aleatoria real no negativa.

En campo de las aplicaciones, el valor de T es el tiempo transcurrido hasta la ocurrencia de un evento (falla). Por ejemplo:

a) En un ensayo clínico, T es el tiempo desde el inicio del tratamiento de un paciente hasta su muerte.
b) En el estudio de enfermedades infecciosas, T representa el tiempo desde el ataque de la infección hasta la aparición de los síntomas.
c) En el estudio de las enfermedades genéticas, T representa desde el nacimiento hasta el inicio de la enfermedad genética, en este caso, T es la edad del individuo.

DEFINICIÓN 1.2.2.- Sea T una variable aleatoria, discreta o continua, la función de distribución acumulada F de T es

$$F(t) = P[T \leq t] \tag{1.1}$$

La **función de densidad** f(t) de la variable aleatoria T, se define por

$$f(t)=\lim_{dt\to 0}\frac{P(t \leq T < t+dt)}{dt},\ 0 < t < \infty \ . \tag{1.2}$$

Donde $f(t)\geq 0$ para todo $t\in[0,T]\subset R$, y tiene un área bajo la curva igual a 1.

Si Fes una función de distribución acumulada, entonces

$$f(t)=\frac{d}{dx}F(t) = F^{'}(t)\ .$$

La función de supervivencia denotado por S(t) se define como:

$$S(t) = P[T > t] = 1 - P[T \leq t] = 1 - F(t) = \int_t^{\infty} f(s)ds \,. \qquad (1.3)$$

En análisis de supervivencia, la función f(t) dada en (1.2) es el límite de la probabilidad de que un individuo fallece en el intervalo [t, t+Δt[por unidad de tiempo. En la figura 1, presentamos las gráficas de las funciones de f(t) y S(t).

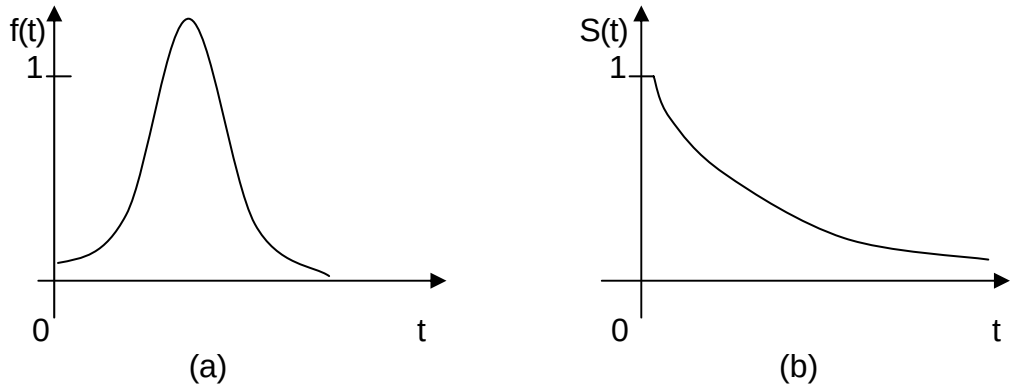

Figura 1: (a) Función de densidad f(t), (b) Función de sobrevivencia S(t) .

Notamos que S(t) es una función continua y monótonamente decreciente, con

$$S(0)=1 \text{ y } S(\infty)=0.$$

La función de supervivencia S(t) se representa como la probabilidad de que un individuo sobreviva más que el tiempo determinado a priori (Lawless, 1982).

PROPOSICIÓN 1.2.3.- Sea S(t) una función de supervivencia. Entonces:

a) S(t)=1 si t<0.

b) $S(\infty)= \lim_{t\to\infty} S(t)=0$.

c) S(t) es una función no-creciente en todo t.

DEMOSTRACIÓN:

a) Sabemos que S(t)= 1-P[T$\leq$t]. Si

$$t<0 \Rightarrow P[T\leq t]=0,$$

porque t es una variable aleatoria no negativa. Luego

$$S(t)=1.$$

b) Para $t=\infty$,

$$S(\infty)=\lim_{t\to\infty} S(t)= \lim_{t\to\infty} (1-F(t))$$

$$= 1-\lim_{t\to\infty} F(t)$$

$$= 1-1=0.$$

c) Por teoría de probabilidades sabemos que F(t) es una función creciente, esto es:

$$t_1 \le t_2 \Rightarrow F(t_1) \le F(t_2), \quad t_1, t_2 \in [0, \tau] \subset R.$$

$$\Rightarrow -F(t_2) \le -F(t_1)$$

$$\Rightarrow 1 - F(t_2) \le 1-F(t_1)$$

$$\Rightarrow \quad S(t_2) \le S(t_1).$$

Por tanto, S(t) es una función no-creciente para $t \in [0, \tau]$. ■

La función S(t) también es conocida como la tasa de sobrevivencia acumulada. Esta función en algunos casos se puede utilizar para determinar el p-ésimo percentil del tiempo de supervivencia. Por ejemplo, el 50-ésimo percentil corresponde al tiempo medio de sobrevivencia.

Observaciones:

a) Si T es una variable aleatoria discreta, entonces

$$f(t) = P[T = t].$$

b) Si T es una variable aleatoria (absolutamente) continua, entonces

$$f(t)= \lim_{\Delta t\to 0} \frac{P(Fallo\, ocurrido\, en [t, t+\Delta t[)}{\Delta t}$$

= razón de ocurrencia de la falla del evento de interés en T=t.

La función de riesgo o función de tasa de fallo, $\lambda(t)$ se define como:

$$\lambda(t)= \lim_{\Delta t \to 0} \frac{P[t \le T < t + \Delta t / T \ge t]}{\Delta t} \qquad (1.4)$$

Esta función representa la tasa instantánea de fallar en el intervalo de tiempo [t, t+Δ t[dado que el individuo ha sobrevivido hasta la edad t=T , (ver figura 2).

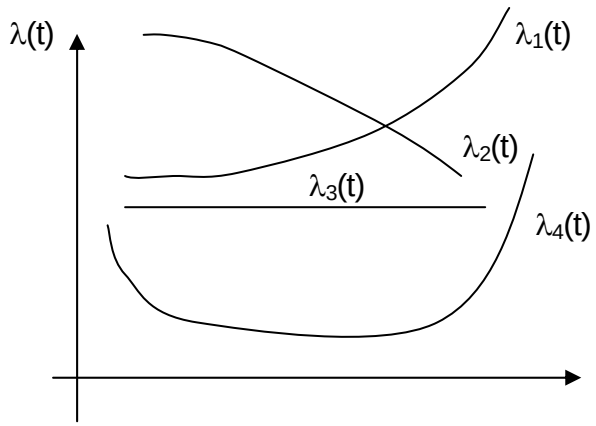

Figura 2: Distintas funciones de riesgo de $\lambda(t)$.

Ejemplo 1:

1. Si $\lambda(t)= \lambda_0>0$, es una función de riesgo constante.
2. Si $\lambda(t)= \lambda_0 + \lambda_1 t$, con $\lambda_0, \lambda_1>0$, es una función de riesgo lineal.

Observación: La expresión

$$\lambda(t)\,\Delta t \cong P[\text{morir entre t y t+}\Delta\text{ t / vivo en t}].$$

Podemos demostrar que

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

En efecto: por definición tenemos que

$$\lambda(t)= \lim_{\Delta t \to 0} \frac{P[t \le T < t + \Delta t / T \ge t]}{\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{P[(t \le T < t + \Delta t) \bigcap (T \ge t)]}{\Delta t P[T \ge t]}$$

$$= \lim_{\Delta t \to 0} \frac{P[t \le T < t + \Delta t]}{\Delta t} \cdot \frac{1}{P[T \ge t]}$$

$$= \frac{f(t)}{S(t)}. \blacksquare$$

Observaciones:

1. Ahora observamos que $\lambda(.)$ esta determinado por f(.) y S(.), o viceversa.
2. El hecho de que

$$\lambda(t)\Delta(t) \simeq P[t \le T < t+\Delta t / T \ge t].$$

Significa

P(expirar en el intervalo [t, t+Δt[/sobrevivió hasta el instante t).

En otras palabras, $\lambda(t)\Delta(t)$ es una probabilidad aproximada de fallar en [t, t+Δt[dado que sobrevivió hasta el instante t.

PROPOSICIÓN 1.2.4.- Las funciones f(t), λ (t) y S(t) cumple las relaciones siguientes:

a) $f(t) = -\frac{d}{dt}(S(t))$.

b) $\lambda(t) = -\frac{d}{dt}(\ln S(t))$.

c) $S(t) = \text{Exp}\left[-\int_0^t \lambda(x)dx \right]$.

Demostración

a) Tenemos que:

$$S(t) = 1 - F(t) \Rightarrow \frac{d}{dt}(S(t)) = -\frac{d}{dt}(F(t))$$

$$\Rightarrow \quad \frac{d}{dt}(S(t)) = -f(t),$$

de donde

$$f(t)=-\frac{d}{dt}(S(t)) . \qquad (1.5)$$

b) Como

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

y, aplicando (1.5) se tiene:

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

$$= - \frac{d}{dt}(\ln S(t)) . \qquad (1.6)$$

c) Por (1.6) sabemos que:

$$\lambda(t)= -\frac{d}{dt}(\ln S(t)) \Rightarrow \lambda(t)dt = -d[\ln S(t)]$$

$$\Rightarrow \int_0^t \lambda(x)dx = -\int_0^t d[\ln S(t)]$$

$$= - \left[\ln S(t)\right]_0^t$$

$$= - \ln S(t) + \ln S(0)$$

$$= - \ln S(t) + \ln(1)$$

$$= - \ln S(t).$$

$$\Rightarrow \ln S(t) = - \int_0^t \lambda(x)dx$$

$$\Rightarrow \ S(t) = \exp\left[-\int_0^t \lambda(x)dx \right]. \qquad (1.7)$$

■

DEFINICION 1.2 5.- Función de riesgo acumulado denotado por $\Lambda(t)$ se define como:

a) Si T es discreto, o sea, los t'¡s son puntos de masa, entonces

$$\Lambda(t)=\sum_{t_i \le t} \lambda(t_i)\,.$$

b) Si T es (absolutamente) continua, entonces

$$\Lambda(t)=\int_0^t \lambda(x)dx\,.$$

La función de riesgo acumulado $\Lambda(t)$ representa la totalidad de las probabilidades de los individuos de fallar en el intervalo [t, t+Δt[dado que han sobrevivido hasta el instante t

Observaciones

1) Si

$$\Lambda(t)=\int_0^t \lambda(x)dx \Rightarrow \frac{d}{dt}(\Lambda(t))=\lambda(t). \qquad (1.8)$$

2) Para alguna variable aleatoria continua T, tenemos que

$$\int_0^t \lambda(x)dx = \infty \Rightarrow e^{-\int_0^t \lambda(x)dx} = e^{-\infty}=0$$

$$\Rightarrow S(\infty)=0.$$

3) En la proposición 1.2.4-(b) aplicando la definición 1.2.5-(b) tendremos que

$$S(t) =e^{-\Lambda(t)}=\exp(-\Lambda(t)). \qquad (1.9)$$

En aplicaciones, si una enfermedad tiene cura; esto es, asumiendo que

$$S(\infty) =P[T=\infty]=q>0.$$

Entonces, por una razón lógica y deductiva debemos tener que

$$\Lambda(\infty)<0 \Rightarrow \int_0^t \lambda(x)dx < \infty. \qquad (1.10)$$

La expresión dada en (1.10) se conoce como el **modelo de cura.** En este caso, T no necesariamente es una variable aleatoria, tal como veremos en el siguiente ejemplo.

Ejemplo 2.- Supongamos que $\lambda(t)=e^{-\theta t}$, $\theta>0$, no es una función de riesgo, puesto que

$$\int_0^\infty \lambda(x)dx = \int_0^\infty e^{-\theta x}dx = \frac{e^{-\theta x}}{-\theta}\Big|_0^\infty$$

$$= -\frac{1}{\theta}[e^{-\theta(\infty)} - e^{-\theta(0)}]$$

$$= -\frac{1}{\theta}[0-1] = \frac{1}{\theta} \neq 0.$$

Pero, si es un modelo de cura, ya que

$$\Lambda(\infty)=\int_0^\infty \lambda(x)dx = \frac{1}{\theta} < \infty.$$

1.3.- DISCRETIZACION DE LOSTIEMPOS DE SUPERVIVENCIA

En muchas ocasiones es deseable tratar a T como una variable aleatoria discreta. Es decir, T asume los valores $t_1,t_2,...t_k$ con

$$0 \leq t_1 \leq t_2 \leq ... \leq t_k$$

y, la función de probabilidad esta dada por:

$$p(t_j)=P[T=t_j],\ j=1,2,\ldots,k.$$

En este caso, la función de supervivencia es

$$S(t)=P[T\geq t]=\sum_{j;t_j\geq t} p(t_j)\,, \qquad (1.11)$$

con S(0)=1 y S(∞)=0. La función de riesgo está dada por

$$\lambda(t_j)=P[T=t_j/T\geq t_j]$$

$$= \frac{P[T=t_j]}{P[T\geq t_j]}$$

$$= \frac{p(t_j)}{S(t_j)}\,,\ j=1,2,...,k \qquad (1.12)$$

Como $p(t_j)=S(t_j)-S(t_{j+1})$ (proporción de fallas en el intervalo $[t_j,t_{j+1}[$) podemos reescribir el (1.12) como

$$\lambda(t_j)=\frac{S(t_j)-S(t_{j+1})}{S(t_j)}$$

$$= 1-\frac{S(t_{j+1})}{S(t_j)}$$

Entonces

$$\lambda(t_j)-1 = -\frac{S(t_{j+1})}{S(t_j)}$$

y

$$1-\lambda(t_j)= \frac{S(t_{j+1})}{S(t_j)}$$

De modo que

$$1-\lambda(t_j)= \frac{P[T\geq t_{j+1}]}{P[T\geq t]}$$

$$= P(T\geq t_{j+1}/\ T\geq t_j),\ j=1,2,...,k. \qquad (1.13)$$

Para

$$t_1\leq t_2\leq \ ...\ \leq t_k\leq t, \qquad (1.14)$$

la función de supervivencia S(t) puede ser escrita como un producto de probabilidades condicionales, ya que, la ocurrencia de los suceso en cada sub-intervalo de (1.14) son eventos independientes , de la siguiente forma.

$$S(t)=P[T\geq t_2/T\geq t_1]P[T\geq t_3/T\geq t_2]\ldots P[T\geq t/T\geq t_k]$$

$$=(1-\lambda(t_1))(1-\lambda(t_2))\ \ldots\ (1-\lambda(t_k))$$

$$=(1-\lambda_1)(1-\lambda_2)\ \ldots\ (1-\lambda_k)$$

$$=\prod_{j:t_j\leq t}(1-\lambda_j)\,. \qquad (1.15)$$

La expresión dada en (1.15) es de mucha utilidad para estimar la función de supervivencia.

Observación: La definición de la función de riesgo acumulado cuando T es discreto es:

$$\Lambda(t)=\sum_{j:t_j\leq t}\lambda_j\ .$$

1.4.-TIPOS DE DATOS CENSURADOS Y ESQUEMA DE CENSURA

Como hemos comentado anteriormente, los datos correspondientes a estudios de Análisis de Supervivencia presentan una particularidad que dificulta su análisis estadístico. Esta particularidad se debe a la presencia de datos censurada, sólo se conoce el tiempo de fallo para una fracción que puede ser pequeña, de los individuos de la muestra; mientras que, del resto sólo se tiene una información parcial de su estado, habitualmente, el tiempo de vida es mayor que un valor dado.

Una observación se dice **censurada por la derecha** en L_1, si se desconoce el valor exacto de la observación y sólo se sabe que ésta es mayor que L_1. Análogamente, una observación se dice **censurado por la izquierda** en L_0, si sólo se sabe que la

observación es menor que el valor L_0. La censura a la derecha es mucho más frecuente que la censura a la izquierda. En algunos experimentos, dependiendo del tipo de problema y el tipo de seguimiento, aparecen datos **censurados en un intervalo** $]t_I , t_D[$; es decir, que sólo se sabe que

$$t_I < T < t_D..$$

Generalmente la duración del tiempo de ensayo se debe limitar por razones prácticas y económicas. Existen dos esquemas básicos para establecer este límite. Más conocido como censuramiento tipo I y tipo II.

1.4.1. CENSURA DE TIPO I.

En este esquema el experimento se programa con una duración C (C>0), es decir, la duración del programa de estudio es desde T=0 hasta $T=t_C$ ($C=t_C-0$) establecida a priori. El tiempo de fallo de un individuo se observará, si es menor o igual que ese valor prefijado t_C. En otro caso, la observación correspondiente será censurado con valor t_C, y la denotaremos con C*, y, el número de observaciones en la muestra es aleatorio.

Sean $T_1,T_2,...,T_n$ los tiempos de supervivencia de los n pacientes de la muestra, los cuales son independientes y idénticamente distribuidos (iid), cada T_i con función de distribución F, en lugar de observar $T_1,T_2,...,T_k$, con $k \leq n$ (k variables aleatorias de interés) podemos solamente observar las variables aleatorias independientes $Y_1,Y_2,...,Y_n$ donde,

$$Y_i = \text{Min}(T_i, t_C) = \begin{cases} T_i & si\ T_i \leq t_C \\ t_C & si\ t_C \leq T_i \end{cases}$$

para i=1,2,...,n, que son los tiempos de supervivencia observadas en la muestra, Gráficamente se tiene.

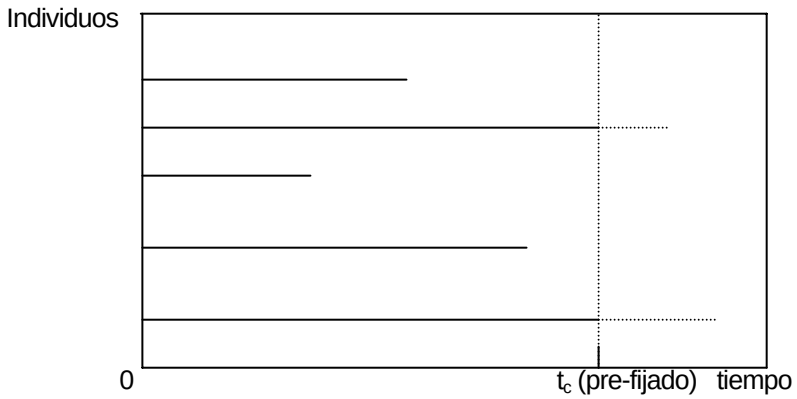

Figura 3: Censura de tipo I

1.4.2. CENSURA DE TIPO II.

En los ensayos realizados bajo un esquema de tipo II, con n componentes idénticos, el ensayo finaliza en el momento en que se produce el r-ésimo fallo ($1 \leq r \leq n$). Ese instante $t_{(r)}$, será el valor de los datos censurados correspondiente a los componentes que en ese momento sigan en estudio. De esta forma sólo se conocen las r observaciones de la muestra y aparecen n-r tiempos censurados en el valor $t_{(r)}=t_C$. Este tipo de censura se usa con frecuencia en los experimentos industriales.

Sean $T_1,T_2,...,T_n$ los tiempos de supervivencia de los n pacientes en la muestra, que son n variables aleatorias (iid) con función de distribución común $F_T(.)$. Sean

$$T_{(1)} \leq T_{(2)} \leq ... \leq T_{(n)}$$

la estadística de orden de $T_1,T_2,...,T_n$. Cesan observaciones después de la r-ésima falla, así podemos observar

$$T_{(1)} ,T_{(2)} ,... , T_{(r)}$$

Que son tiempos de falla y, la muestra completa observada es:

$$\left.\begin{array}{c} Y_1 = T_{(1)} \\ \vdots \\ Y_r = T_{(r)} \end{array}\right\} \text{(no censurados)}$$

$$\left.\begin{array}{c} Y_{r+1} = T_{(r)} \\ \vdots \\ Y_n = T_{(r)} \end{array}\right\} \text{(censurados)}$$

Gráficamente tendremos:

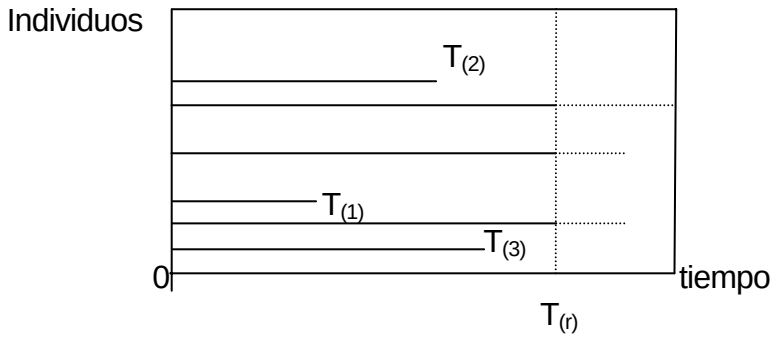

Figura 4: Censura de tipo II

Es importante señalar que el valor C en el esquema de tipo I y el valor de r (o la fracción r/n) que indica la tasa de censura en el esquema de tipo II debe fijarse antes de iniciar el experimente y no durante el transcurso del mismo dependiendo de los resultados que se observen. La necesidad de que el mecanismo de censura sea independiente de la observación del fenómeno, es un requisito imprescindible para la validez de las conclusiones.

1.4.3. CENSURAMIENTO ALEATORIO.

En el campo de la Fiabilidad que corresponde al ámbito industrial y tecnológico, llamamos componente o sistema a un conjunto de elementos que funcionan bajo una tensión mecánica o eléctrica. El sistema dejará de funcionar cuando fallan r elementos y, el tiempo correspondientes a este suceso final es $t_{(r)}$ ó C; y los valores obtenidos como variable respuesta T se llaman **muestra simplemente**

censurados, es decir, tienen con un único valor común $t_{(r)}$ ó C**,** de las observaciones censuradas. En los ensayos médicos, aunque el experimento se diseñe con una limitación temporal como en el esquema I, es normal que los individuos se incorporen al ensayo en instantes aleatorios, y también es habitual que se produzcan abandonos durante la realización del ensayo o estudio. Posteriormente ese mismo individuo puede regresar para incorporarse al estudio. En consecuencia, las muestras resultantes en este caso pueden ser *varias veces censuradas*. Generalmente, este tipo de observaciones se presentan mediante un par de variables (T,δ), donde, T es el tiempo transcurrido desde la entrada del individuo al ensayo hasta la salida del mismo y δ es una variable binaria indicadora del tipo de observación, que toma el valor 1 si se ha observado el fallo y el valor 0 si se trata de una observación censurado.

De manera más formal se tiene. Sean $T_1,T_2,...,T_n$ los tiempos de supervivencia de los n individuos en la muestra, que son variables aleatoria (iid) con función de distribución común F_T, tal que $F_T(0)=0$. Sean $C_1, C_2,...,C_n$ los tiempos de

censuramiento (el periodo de observación límite) correspondiente a los n pacientes, que se asume que son variables aleatorias (iid) con función de distribución común G_C, generalmente no conocida. También asumiremos que los T_k y C_k son independientes, de modo que se observa en la muestra los tiempos

$$Y_1, Y_2, ..., Y_n$$

donde,

$$Y_k=\min(T_k,C_k),\ k=1,2,...,n.$$

Y, sea δ_k la función indicadora de censuramiento de la k-ésima observación definida por:

$$\delta_k= I_{[T_k \le C_k]} = \begin{cases} 1 \ si \ \ Y_k \ es\ no\ censurado \\ 0 \ si \ Y_k \ \ es\ censurado \end{cases},\ k=1,2,...,n$$

en consecuencia, los valores observados en la muestra serán los pares

$$(Y_1,\delta_1), (Y_2,\delta_2),....,(Y_n,\delta_n)$$

donde los Y'_is son variables aleatorias *iid*, con distribución común H(.,.), y

los $\delta_1, \delta_2,..., \delta_n$ contienen la información de censuramiento. Gráficamente podemos representar por :

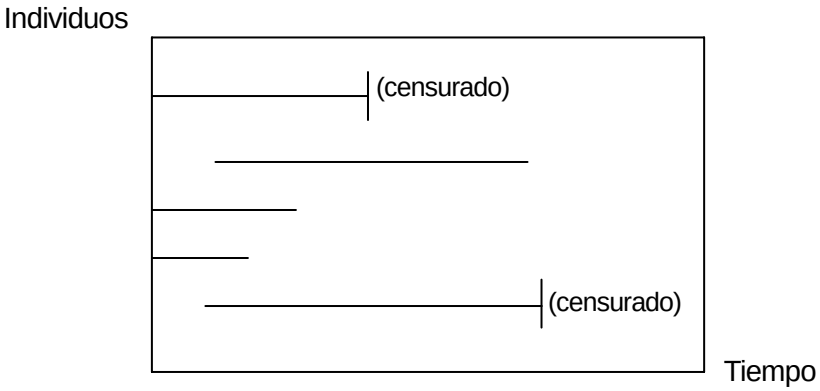

Figura 5: Censura Aleatorio

Ejemplo 3: Podemos notar que, en las muestras no solamente pueden ser censuradas varias veces, sino también, pueden ser observadas varias veces los eventos de interés en los individuos. Este tipo de observaciones repetitivas en los individuos se llama **eventos recurrentes**. Sea un conjunto de datos de sobrevivencia:

$$25, 18^{+}, 17, 22^{+}, 27.$$

Los cuales pueden ser representados por:

T_i	25	18	17	22	27
δ_i	1	0	1	0	1

En análisis de supervivencia, algunas causas de censuramiento pueden ser:

- Pérdida de seguimiento.
- Abandonos.
- Muerte por otras razones diferentes de la de interés.
- Desconfianza.
- Final programado del estudio para el análisis, etc, etc.

Como uno de los objetivos del presente trabajo de tesis es la utilización del Lenguaje **R** en el tratamiento sistemático de los datos de supervivencia, tal como se introduce en el siguiente ejemplo. Para tal efecto, utilizaremos los datos siguientes que muestran los tiempos de supervivencia, dados en meses, de 24 pacientes de cáncer de colon (Rivas y López, Análisis de Supervivencia). Los valores de tiempo seguido de un signo + indican que el paciente se encuentra en condición de censurado.

3+, 6, 6, 6, 6, 8, 8, 12, 12, 12+, 15+, 16+,

18+, 18+, 20, 22+, 24, 28+, 28+, 28+, 30, 30+, 33+, 42

Para alcanzar nuestro objetivo propuesto con datos censurados, utilizaremos siempre la instrucción **library (survival)** del lenguaje **R.**

> library(survival)

y los datos se representan por:

> tiempo<-c(3,6,6,6,6,8,8,12,12,12,15,16,18,18,20,22,24,28,28,28,30,30,33,42)

> estado<-c(0,1,1,1,1,1,1,1,1,0,0,0,0,0,1,0,1,0,0,0,1,0,0,1)

y la parte operativa se hace con:

> Surv(tiempo,estado)

[1] 3+ 6 6 6 6 8 8 12 12 12+ 15+ 16+ 18+ 18+ 20 22+ 24 28+ 28+

[20] 28+ 30 30+ 33+ 42

CAPITULO II

ESTIMACION NO PARAMETRICA.

2.1.- INTRODUCCIÓN.

El problema principal en los problemas de Análisis de Supervivencia es la estimación de la función de supervivencia S(t). Esta función es la base para estimar la mayor parte de las funciones y parámetros de interés en el análisis del tiempo de vida.

2.2.- ESTIMACIÓN NO PARAMÉTRICA DE LA SUPERVIVENCIA.

Si la muestra no contiene observaciones censuradas, la función de supervivencia se estima mediante **función de supervivencia empírica, (fse)** definida como,

$$\hat{S}(t) = S_n(t) = \frac{N^0 \text{ de individuos con tiempo de supervivencia mayor que } t}{N^0 \text{ de individuos de la muestra}} \qquad (2.1)$$

Para $t \geq 0$. El símbolo $\wedge$ denota el estimador de la función S.

Si miramos a $\hat{S}(t)$ como una función de t, este estimador es una función no creciente, que toma el valor 1 en todo instante anterior al tiempo de muerte (falla) más pequeño y, 0 a partir del máximo tiempo de muerte observada; la

función permanece constante entre dos instantes de muerte consecutivos. Si no hay empates en la muestra, todos los saltos descendentes de la función son iguales a 1/n, mientras que si se observa d tiempos de vida iguales (empates) a t_i , i=1,2,...k el salto descendente de $\hat{S}(t)$ en ese instante será igual a d/n. Tal como se muestra en la figura siguiente.

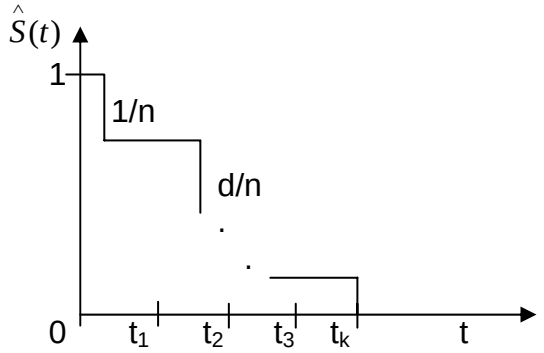

Figura 6: Función de sobrevivencia estimada $\hat{S}(t)$

Cuando en la muestra existen observaciones censuradas la función de sobrevivencia dada en (2.1)**,** no es un estimador adecuado porque tiende a subestimar la función de supervivencia. En efecto, el utilizar este estimador es equivalente a considerar que todos los individuos censurados fallan en el instante de censura. Dado que es posible que alguno de los individuos con tiempo de censura menor que t esté vivo en el instante t, por tanto, será necesario introducir alguna modificación en el estimador para evitar el sesgo.

Supongamos que en una muestra de n individuos, se han observado m muertes diferentes, entonces, para m≤ n, si tiene que n-m≥0, que es el número de pacientes que pasaron el tiempo t, luego según (2.1) se tiene:

$$\hat{S}(t) = \frac{n-m}{n} = 1 - \frac{m}{n}. \qquad (2.2)$$

Observación: La fórmula (2.2) se aplica siempre en cuando que no hay elementos censurados.

En el estudio de los elementos censurados podemos notar algunos apreciaciones importantes, tales como:

a) El tiempo de observación [0, T] se subdivide en partes t_1, t_2,...,t_k, y que, sea n_j el número de individuos vivos justos antes de t_j. La probabilidad de que un individuo muera entre t_{j-1} y t_j, si estaba vivo en el instante t_{j-1} es $1/n_j$. Por tanto, la probabilidad de que un individuo que vivía en t_{j-1} sobreviva el momento t_j es

$$1- \frac{1}{n_j}.$$

b) Si murieran m_j individuos entre t_{j-1}, t_j sabiendo que estaban vivos en tj_{-1}, en este caso la probabilidad de morir es

$$\frac{m_j}{n_j},$$

y,

$$\frac{n_j - m_j}{n_j} = 1- \frac{m_j}{n_j}, \; j=1,2,....k \qquad (2.3)$$

es la probabilidad de que $n_j - m_j$ individuos sobrevivan en el momento t_j.

Ejemplo 4:- Supongamos que una dosis de una cierta droga es aplicada a un grupo de gatos, los tiempos de sobrevida observada son:

dosis	n	Tiempos de vida (t_i)
2.5	10	5 6 6 7 8 8 8 10 11 12

Estimar S(t) y hacer el gráfico respectivo.

Solución:

Tenemos:

$$\hat{S}(5)=S_{10}(5)=1-\frac{1}{10}=0.9$$

$$\hat{S}(6)=S_{10}(6)=1-\frac{3}{10}=0.7$$

$$\hat{S}(7)=S_{10}(7)=1-\frac{4}{10}=0.6$$

$$\hat{S}(8)=S_{10}(8)=1-\frac{7}{10}=0.3$$

$$\hat{S}(10)=S_{10}(10)=1-\frac{8}{10}=0.2$$

$$\hat{S}(11)=S_{10}(11)=1-\frac{9}{10}=0.1$$

$$\hat{S}(12)=S_{10}(12)=1-\frac{10}{10}=0.0$$

Los resultados anteriores podemos representarlos en una tabla.

Tabla 1: Valores estimados de $\hat{S}(t)$ según los valores empatados d del ejemplo.

t	$\hat{S}(t)$	d
5	0.9	1
6	0.7	2
7	0.6	1
8	0.3	3
10	0.2	1
11	0.1	1
12	0.0	1

Gráficamente se tiene la figura 7:

La función de densidad de probabilidad f(t) puede ser estimada a partir de los datos muestrales por medio de la expresión siguiente:

$$\hat{f}(t)=\frac{N^0\ de\ pacientes\ que\ fallecieron\ en\ el\ \text{int}ervalo\ comenzando\ en\ t}{(N^0\ total\ de\ pacientes)(Amplitud\ del\ \text{int}ervalo)}, \qquad (2.4)$$

y, la función de riesgo h(t) es estimada por:

$$\hat{\lambda}(t) = \frac{N^0 \text{ de pacientes que fallecieron en el intervalo iniciando en } t}{(N^0 \text{ de pacientes con tiempos} > t)(\text{Amplitud del intervalo})}, \qquad (2.5)$$

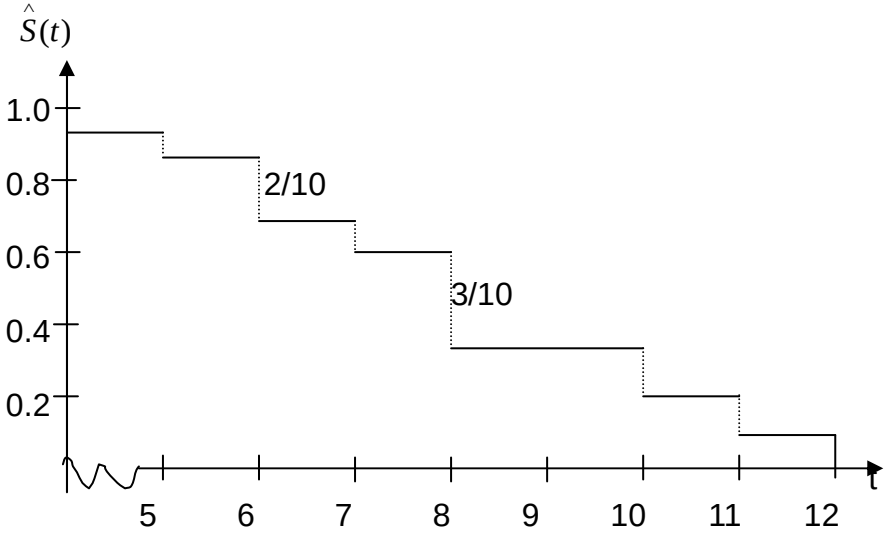

Figura 7: Gráfica de $\hat{S}(t)$ según la tabla 2.1

.

Ejemplo 5.- Francisco Louzada Neto y otros, en su libro Análisis de Sobrevivencia y Confiabilidad-2002. En la pág. 20 presenta los resultados de tiempos de sobrevivencia (en dias) de pacientes que recibieron el transplante de corazón.

2 3 7 7 10 10 17 21 26 26 28 43 44 50
52 56 57 59 62 62 68 68 69 78 79 88 91 97
98 98 104 106 119 135 143 150 154 162 190 209 228 242
248 249 252 273 291 297 311 334 340 346 354 366 391 421
439 470 478 481 490 495 570 583 614 615 621 652 697 730
773 776 790 793 806 840 852 875 939 945 945 946
1013 1105 1164 1186 1191 1210 1326 1331 1357 1388 1418 1473 1509
1734 1777 1820 1835 1877 1940 2034 2056 2108 2291 2301 2313
2421 2489 2567 2795 3146.

Como podemos observar en este conjunto, no hay datos censurados. Para estimar los valores de f(t), S(t) y λ(t) utilizaremos los comandos del software R. Por lo que a x se asigna los datos anteriores mediante la función c(.).

> x<- c(2, 3, 7, 7,…, 2795, 3146)

> datos<-x

> int<-hist(datos,plot=F)

> tab<-table(cut(datos,int$breaks))

> S<-S[S>0]

> λ<-f/S

> tabla<-cbind(as.matrix(tab),f,S,h)
> tabla.

Tabla 2: Valores de la función densidad, función de sobrevivencia y de riesgo estimada.

	frec	f	S	λ
(0,500]	62	1.107143e-03	1.000000000	0.0011071426
(500, 1e+03]	20	3.571429e-04	0.446428571	0.0008000000
(1e+03, 1.5e+03]	12	2.142857e-04	0.267857143	0.0008000000
(1.5e+03, 2e+03]	7	1.250000e-04	0.160714286	0.0007777778
(2e+03, 2.5e+03]	8	1.428571e-04	0.098214286	0.0014545455
(2.5e+03, 3e+03]	2	3.571429e-05	0.026785714	0.0013333333
(3e+03, 3.5e+03]	1	1.785714e-05	0.008928571	0.0020000000

las frecuencias 62, 20, 12, 7, 8, 2 y 1 en total suman 112, representan a los individuos que fallecen en los correspondientes intervalos de tiempo. Las respectivas gráficas se obtienen de la manera siguiente:

Gráfica de la función de densidad f(t) y de la función de supervivencia estimada S(t).
:

>plot(int$mids,int$density,type="b",xlab="Tiempo",ylab="Función Densidad")

>plot(int$mids,S,type="b",xlab="Tiempo",ylab="Función Supervivencia")

Obteniendo:

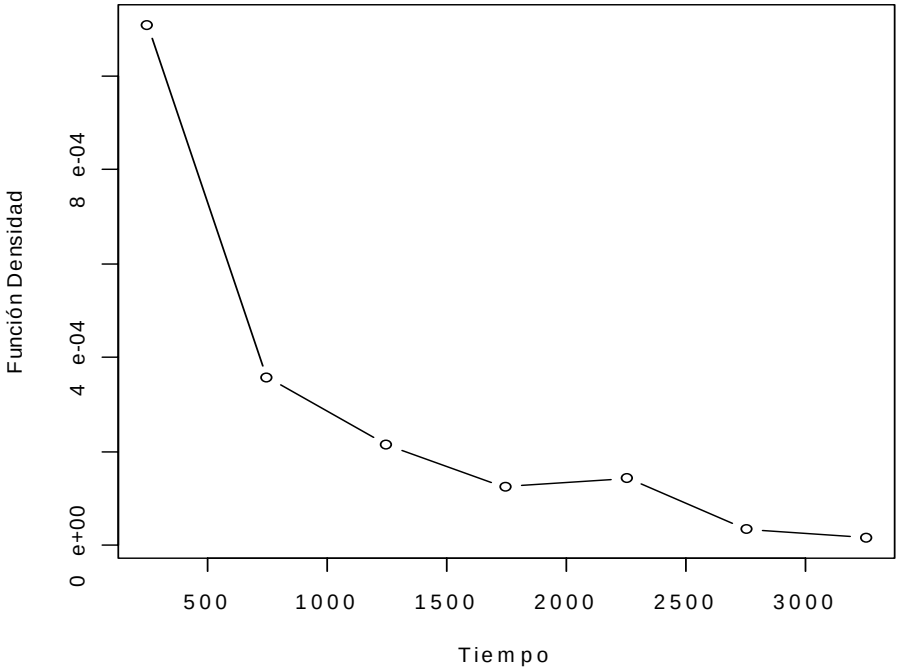

Figura 8: Función de densidad f(t) de la tabla 2, obtenido según el lenguaje R.

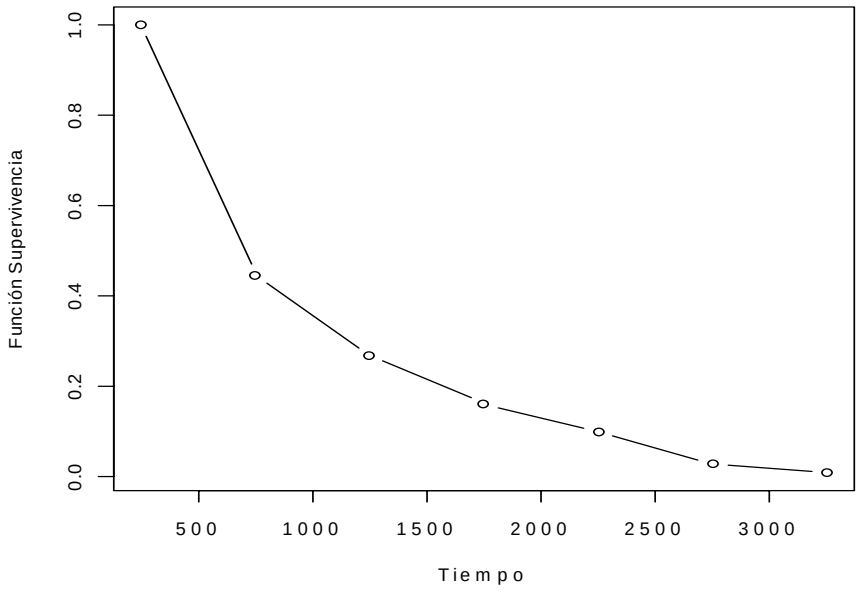

Figura 9: Función de supervivencia S(t) de la tabla 2, obtenido Según el lenguaje R.

Gráfica de la función de riesgo estimada h(t).

> plot(int$mids,λ,type="b",xlab="Tiempo",ylab="Función Riesgo").

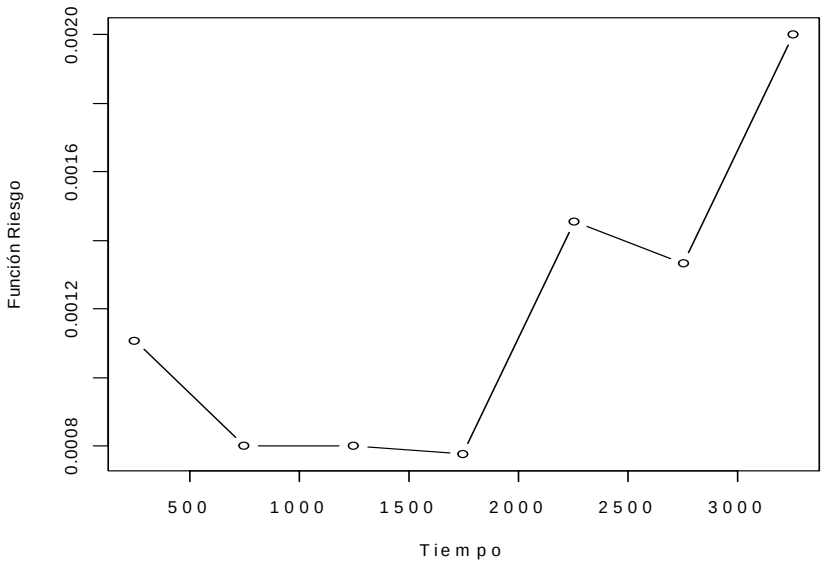

Figura 10: Función de riesgo λ(t) de la tabla 2.

2.3.- EL ESTIMADOR DE KAPLAN-MEIER.

En la sección anterior describimos como estimar con simplicidad las funciones de densidad de probabilidad, de sobrevivencia y de riesgo, a partir de un conjunto de datos muestrales y según las expresiones (2.1), (2.4) y (2.5). Sin embargo, estas expresiones no permiten la presencia de observaciones censuradas, las cuales son comunes en los datos de supervivencia y confiabilidad. En esta sección describiremos como las funciones de supervivencia y de riesgo pueden ser estimadas por medio del **estimador de Kaplan-Meier** (1958), el cual permite la presencia de observaciones censuradas. Este estimador es también conocido en la literatura como el **estimador de producto-límite** por disponer de los tiempos de supervivencia individuales (Louzada Neto y Otros-2002).

Este estimador no-paramétrico es más utilizado en la estimación de la función de supervivencia S(t), que se basa en el principio de calcular la supervivencia como producto de probabilidades condicionadas, según (1.15) cuyo resultado es:

$$S(t)=\prod_{j:t_j<t}(1-\lambda(t_j)) \qquad (2.6)$$

pero, llevando la partición del tiempo de estudio en intervalos al caso extremo de considerar que cada intervalo contenga sólo la observación correspondiente a un individuo, sea ésta fallo o censura; de ahí, viene el nombre de producto límite. Si los datos observados corresponden a los tiempos

$$0=t_0<t_1<...<t_n$$

se considera la partición determinada por los intervalos

$$]t_{i-1}, t_i] \ , \ i=1,2,...,n$$

al que pertenece el instante de tiempo t_i , pero no al t_{i-1} y además el interior de estos intervalos está siempre libre de censura, que sólo ocurrirán en su caso, en un extremo del subintervalo. Si llegan n_i individuos con vida al intervalo

$$]t_{i-1}, t_i] \ , \ i=1,2,...,n$$

el estimador de la probabilidad de fallo en ese intervalo, condicionado a haber sobrevivido hasta entonces, se define por:

$$q_i=\begin{cases}\dfrac{1}{n_i} & si\ \ en\ \ t_i\ \ se\ \ produce\ un\ \ fallo\\ 0\ \ si\ en\ t_i\ \ se\ \ produce\ una\ censura\end{cases}$$

para i=1,2,....n.

Entonces, la probabilidad de que un individuo que vivía en t_{i-1} sobreviva el momento t_i es:

$$p_i=1-q_i=\begin{cases}1-\dfrac{1}{n_i} & si\ en\ t_i\ \ se\ \ produce\ un\ \ \ fallo\\ 1\ \ si\ \ en\ \ t_i\ \ se\ \ produce\ una\ \ censura\end{cases}$$

para i=1,2,...,n. O sea,

$$p_i=P(\text{sobrevivir hasta } t_i/ \text{ sobrevivió en } t_{i-1}).$$

Por otro lado, los intervalos que no contienen los fallos no contribuyen a la construcción de S(t), ya que, para ellos la estimación de la probabilidad condicionada de supervivencia en el intervalo es 1. Sin embargo, la existencia de censura si influye en el número de individuos expuestos al riesgo de fallar al comienzo del intervalo siguiente, que se ve disminuido en una unidad.

Es posible que en la muestra se produzca empates, es decir, observaciones cuyo tiempo de fallo es el mismo y por eso se define d_i ($d_i \geq 1$) como el número de fallos que se producen en el instante t_i , i=1,2,...,n. Para mayor claridad utilizaré las siguientes designaciones:

- n_i : Número de individuos que llegan al comienzo del intervalo (número de individuos en riesgo al inicio del intervalo).
- d_i : Número de fallecidos in el intervalo $]t_i, t_i+1]$.
- p_t : Proporción de individuos que han sobrevivido al instante t. Esto es:

$$p_t = \frac{n_t - d_t}{n_t} = 1 - \frac{d_t}{n_t}$$

que representa, la probabilidad condicional de sobrevivir el j-ésimo tiempo, habiendo sobrevivido hasta el instante (j-1)-ésimo.

Observación Si un tiempo de vida y uno de censura coinciden, se considera como individuo en riesgo, esto es, cualquier individuo con tiempo de censura igual a t_j se incluye en el conjunto de n_j individuos en riesgo en t_j.

2.3.1 ESTIMACIÓN DE FUNCIÓN DE RIESGO λ(t).

Como sabemos que

$$1-\lambda\ (t_j),$$

es la probabilidad condicional de sobrevivir al instante t_j dado que sobrevivió hasta el t_{j-1}, entonces por analogía con (2.1) se tiene que

$$1-\hat{\lambda}(t_j) = \frac{numero\ de\ observaciones > t}{n_j}$$

$$= \frac{n_j - d_j}{n_j}$$

$$= 1 - \frac{d_j}{n_j} = \hat{p}_j \,. \qquad (2.7)$$

Ahora, utilizando el resultado (2.7) en (2.6), el estimador de la función de supervivencia es

$$\hat{S}(t) = \prod_{j:t_j<t} (1 - \frac{d_j}{n_j}) = \prod_{j:t_j<t} \hat{p}_j \,, \qquad (2.8)$$

que se conoce como el **estimador de Kaplan-Meier.**

Observación: Cuando t=0, $\hat{S}(t)$=1; es decir, todos los individuos comienzan vivos el estudio.

2.3.2.-DETERMINACIÓN DE LA VARIANZA DE $\hat{S}(t)$.

Aplicamos logaritmo a (2.8); o sea,

$$\log(\hat{S}(t)) = \sum_{j:t_j<t} \log(\hat{p}_j) \,. \qquad (2.9)$$

Desarrollamos en una expansión de series de Taylor el $\log(\hat{p}_j)$ alrededor de $\hat{p}_j = p_j$. Así:

$$\log(\hat{p}_j) = \log(p_j) + \frac{d}{dp_j}\Big(\log(p_j)\Big)(\hat{p}_j - p_j) + \frac{1}{2}.\frac{d^2}{dp_j^2}\Big(\log(p_j)\Big)(\hat{p}_j - p_j) + R(\hat{p}_j, p_j)$$

para j=1,2,...,n. De donde aproximamos

$$\log(\hat{p}_j) \simeq \frac{d}{dp_j}\Big(\log(p_j)\Big)(\hat{p}_j - p_j), \ j=1,2,...,n.$$

Ahora aplicamos varianza a ambos lados:

$$\text{Var}(\log(\hat{p}_j)) \simeq \left[\frac{d}{dp_j}(\log(p_j)\right]^2 Var(\hat{p}_j - p_j)$$

$$= \left[\frac{d}{dp_j}(\log(p_j)\right]^2 [Var(\hat{p}_j) + \text{Var}(p_j)],$$

es decir,

$$\text{Var}(\log(\hat{p}_j)) \simeq Var(\hat{p}_j)\left[\frac{d}{dp_j}(\log(p_j)\right]^2, \quad j=1,2,...,n. \qquad (2.10)$$

Por otro lado, asumimos que

$$n_j \hat{p}_j \sim \text{Binomial}(n_j, p_j), \quad j=1,2,...,n$$

entonces

$$\text{Var}(n_j \hat{p}_j) = n_j p_j(1-p_j) \Rightarrow n_j^2 \text{Var}(\hat{p}_j) = n_j p_j(1-p_j)$$

$$\Rightarrow \text{Var}(\hat{p}_j) = \frac{p_j(1-p_j)}{n_j}; \quad j=1,2,...,n.$$

Luego en (2.10) se tiene:

$$\text{Var}(\log(\hat{p}_j)) \simeq \frac{p_j(1-p_j)}{n_j} \cdot \frac{1}{p_j^2}, \quad j=1,2,..,n$$

$$= \frac{1-p_j}{p_j n_j}, \quad j=1,2,...,n \qquad (2.11)$$

y, supongamos en (2.9) que

$$\log \hat{p}_1, \log \hat{p}_2, ..., \log \hat{p}_j$$

son independientes, entonces en (23) se tiene

$$\text{Var}[\log(\hat{S}(t))] = \sum_{j:t_j<t}\left[\frac{1-\hat{p}_j}{\hat{p}_j n_j}\right]$$

$$= \sum_{j:t_j<t}\left[\frac{1-(1-\frac{d_j}{n_j})}{(1-\frac{d_j}{n_j})n_j}\right] = \sum_{j:t_j<t}\left[\frac{\frac{d_j}{n_j}}{n_j - d_j}\right],$$

de modo que

$$\text{Var}[\log(\hat{S}(t))] = \sum_{j:t_j<t}\left[\frac{d_j}{n_j(n_j-d_j)}\right]. \qquad (2.12)$$

De manera análoga, aplicaremos el desarrollo de la serie de Taylor al $\log \hat{S}(t)$ alrededor de $\hat{S}(t)$ =S(t) tendremos:

$$\log \hat{S}(t) = \log S(t) + \frac{d}{dS(t)}[\log S(t)](\hat{S}(t) - S(t)) + R(\hat{S}(t), S(t)).$$

De donde:

$$\log \hat{S}(t) \simeq \frac{d}{dS(t)}[\log S(t)](\hat{S}(t) - S(t))$$

$$= \frac{1}{S(t)}(\hat{S}(t) - S(t)),$$

y

$$\text{Var}(\log \hat{S}(t)) \simeq \frac{1}{S^2(t)} Var(\hat{S}(t)),$$

reemplazando este resultado en (2.13) se tiene:

$$\frac{1}{S^2(t)} Var(\hat{S}(t)) \simeq \sum_{j:t_j<t}\left[\frac{d_j}{n_j(n_j-d_j)}\right]$$

finalmente, obtendremos

$$Var(\hat{S}(t)) \simeq \hat{S}^2(t) \sum_{j:t_j<t}\left[\frac{d_j}{n_j(n_j-d_j)}\right]. \qquad (2.13)$$

De donde el error estándar de S(t) se puede estimar por

$$\text{e.e}(\hat{S}(t)) \simeq \hat{S}(t)\left[\sum_{j:t_j<t}\frac{d_j}{n_j(n_j-d_j)}\right]^{1/2} \qquad (2.14)$$

y, el intervalo de confianza al 95% se calcula de la manera usual

$$S(t) \pm 1.96\ \text{e,e}(\hat{S}(t)),$$

esto nos permite observar las bandas de confianza para S(t). Gráficamente se tiene.

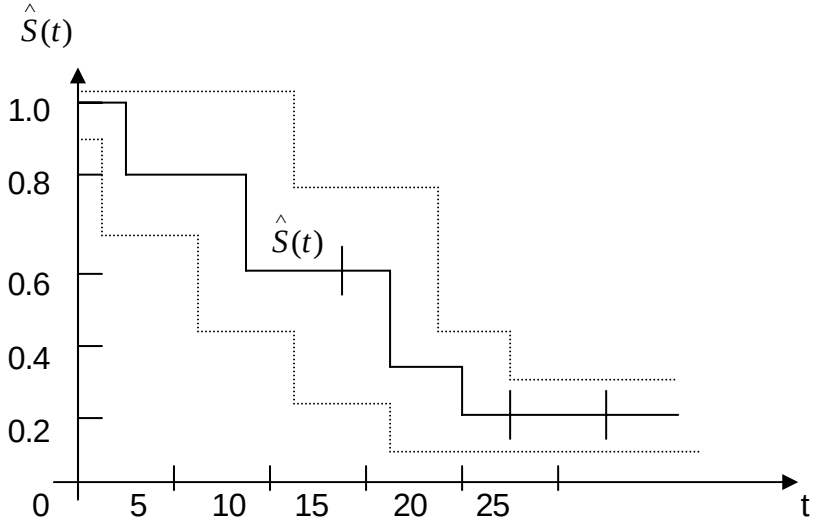

Figura 11: Bandas de confianza para $\hat{S}(t)$

Ejemplo 6: Gehan estudió los resultados de un ensayo clínico sobre la eficacia de la droga 6-mercaptopurina (6-MP) para prolongar el estado de remisión, es decir, la ausencia de síntomas de la enfermedad, en enfermos que habían padecido leucemia aguda. Con el fin de contrastar su efecto, se administró esta droga a un grupo de pacientes, mientras a otro grupo, que servía de control, se le administró un placebo. La asignación de pacientes a cada uno de los dos grupos se hizo aleatoriamente. En el enyaso participaron 42 individuos cuyos tiempos de remisión, registrados en semanas, se muestra en la tabla 3.

Tabla 3: Eficacia de 6-MP y del grupo control (placebo).

Placebo	1, 2, 2, 2, 3, 4, 4, 5,5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23
6-MP	6, 6, 6, 6+, 7, 9+, 10, 10+, 11+, 13, 16, 17+, 19+, 20+, 22, 23 25+, 32+, 32+, 34+, 35+

El periodo de observación fue de un año y, durante ese intervalo, los enfermos se fueron incorporando al ensayo más o menos regularmente. Las observaciones seguidas “+” son censurados. En esos pacientes, la enfermedad

estaba aún en estado de remisión en el momento en que se les efectuó el último control.

PLACEBO:

>library(survival)

> tiempo<-c(1,2,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23)

> estado<-c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)

> summary(survfit(Surv(tiempo,estado)))

Call: survfit(formula = Surv(tiempo, estado))

Al ejecutar éstas instrucciones se obtiene la tabla 4

Tabla 4: Valores de la función de supervivencia del grupo placebo según el lenguaje R.

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	21	1	0.9524	0.0465	0.86552	1.000
2	20	3	0.8095	0.0857	0.65785	0.996
3	17	1	0.7619	0.0929	0.59988	0.968
4	16	2	0.6667	0.1029	0.49268	0.902
5	14	2	0.5714	0.1080	0.39455	0.828
8	12	4	0.3810	0.1060	0.22085	0.657
11	8	2	0.2857	0.0986	0.14529	0.562
12	6	2	0.1905	0.0857	0.07887	0.460
15	4	1	0.1429	0.0764	0.05011	0.407
17	3	1	0.0952	0.0641	0.02549	0.356
22	2	1	0.0476	0.0465	0.00703	0.322
23	1	1	0.0000	NA	NA	NA

6-MP:

> tiempo1<-c(6,6,6,6,7,9,10,10,11,13,16,17,19,20,22,23,25,32,32,34,35)
> estado1<-c(1,1,1,0,1,0,1,0,0,1,1,0,0,0,1,1,0,0,0,0,0)

> summary(survfit(Surv(tiempo1,estado1)))

Tabla 5. Valores de sobrevivencia estimada

.

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
6	21	3	0.857	0.0764	0.720	1.000
7	17	1	0.807	0.0869	0.653	0.996
10	15	1	0.753	0.0963	0.586	0.968
13	12	1	0.690	0.1068	0.510	0.935
16	11	1	0.627	0.1141	0.439	0.896
22	7	1	0.538	0.1282	0.337	0.858
23	6	1	0.448	0.1346	0.249	0.807

Para obtener la gráfica correspondiente al control (placebo) utilizaremos los siguientes comandos del software R.

> fit1<-survfit(Surv(tiempo,estado))

> plot(fit1,xlab="tiempo",ylab="función supervivencia")

Para obtener la gráfica correspondiente al tratamiento 6-MP.

> fit2<-survfit(Surv(tiempo1,estado1))

> plot(fit2,xlab="tiempo",ylab="función supervivencia")

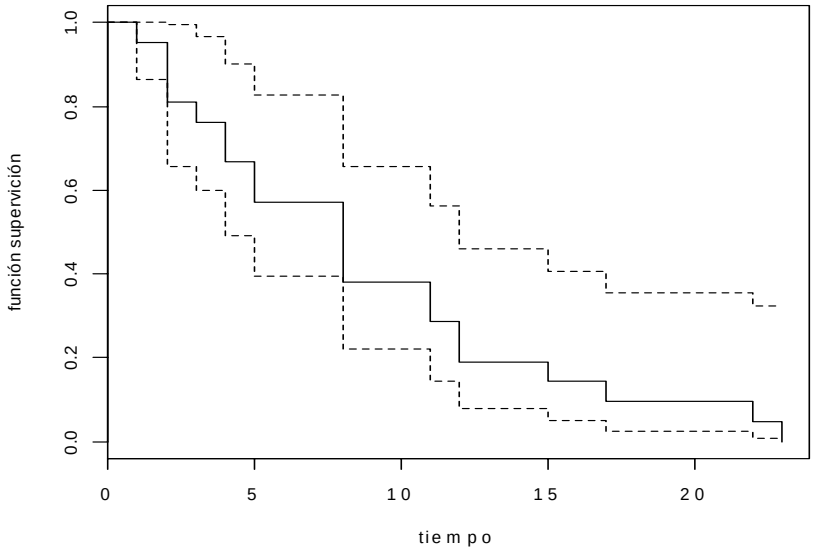

Figura 12: Gráfica de los valores de $\hat{S}(t)$ del placebo.

En las figuras 12 y 13 presentamos las gráficas de las funciones de supervivencia estimadas por el método de Kaplan-Meier, para los dos tipos de enyaso clínico. Notamos que las curvas estimadas es de tipo escalonada, la primera no presenta datos censurados,

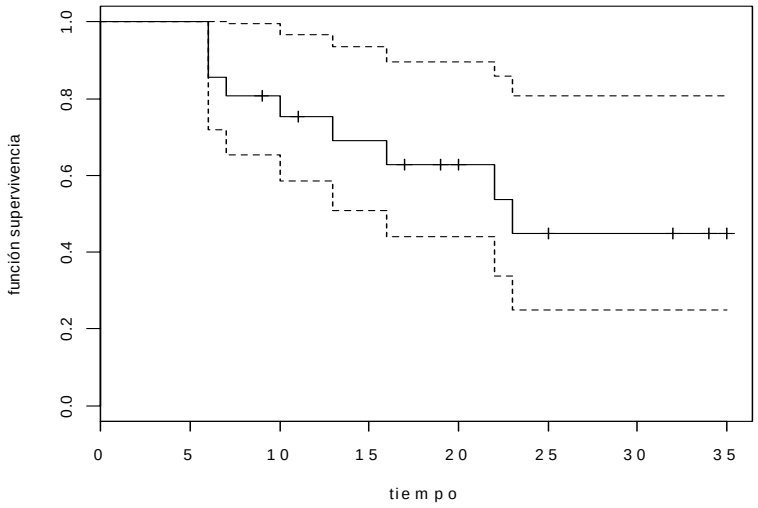

Figura 13: Gráfica de los valores de $\hat{S}(t)$ Del grupo 6-MP.

mientras que la segunda gráfica si presenta datos censurados indicando con “+”, es más, que las probabilidades de supervivencia son constantes entre los tiempos de muerte adyacentes y, más alargada en cada tiempo distinto a medida que el número de la muestra decrece. Observamos que los individuos partícipes en el grupo de control (placebo) tiene una sobrevivencia menor que las personas tratadas con 6-MP.

2.4.- COMPARACIÓN DE DOS FUNCIONES DE SUPERVIVENCIA.

Un problema frecuente en los estudios de Análisis de Supervivencia es la comparación de dos poblaciones que se diferencian en alguna característica, por ejemplo, el tratamiento aplicado.

2.4.1.-PRUEBA LOG-RANK PARA DOS MUESTRAS.

Supongamos que se quiere comparar la supervivencia en dos grupos de individuos, A y B, y que se dispone de una muestra de cada población de tamaño n_A y n_B respectivamente; sea

$$n_A + n_B$$

el tamaño de la muestra combinada y,

$$t_{(1)}<...<t_{(j)}<\ldots<t_{(J)},$$

los J tiempos de fallos distintos observados en la muestra conjunta ordenada en forma creciente. Denotaremos por d_{A_j} y d_{B_j}, el número de fallos ocurridos en el instante $t_{(j)}$ en cada grupo y por d_j el número total de fallos observados en ese instante, es decir,

$$d_j = d_{A_j} + d_{B_j} .$$

Análogamente, denotaremos por n_{A_j} y n_{B_j} el número de individuos en riesgo en cada grupo justo antes del instante $t_{(j)}$, y por n_j la suma

$$n_{A_j} + n_{B_j} .$$

Toda esta información se puede disponer en un conjunto de J tablas de contingencia 2x2, una para cada tiempo de fallo. La tabla correspondiente al instante $t_{(j)}$ se tiene.

Tabla 2.6: Valores observados en los grupos A x B.

Grupo	N° fallos en $t_{(j)}$	N° individuos vivos en $t_{(j)}$	N° indiviuos en riesgo en $t_{(j)}$
A	d_{A_j}	n_{A_j} - d_{A_j}	n_{A_j}
B	d_{B_j}	n_{B_j} - d_{B_j}	n_{B_j}
Total	d_j	$n_j - d_j$	n_j

La hipótesis a contrastar es que las funciones se supervivencia en ambos grupos coinciden en el intervalo de tiempo observado; esto es,

$$H_o : S_A(t) = S_B(t), \quad t \leq \tau.$$

O, equivalentemente

$$H_o : \lambda_A(t) = \lambda_B(t), \quad t \leq \tau,$$

donde, en general, τ se toma igual al mayor tiempo de supervivencia observada. Para contrastar esta hipótesis consideraremos la discrepancia que se observa entre los individuos que fallan en cada uno de los grupos en cada instante de fallo, y el correspondiente número esperado de fallos bajo H_o. Dada la historia de fallo y censura hasta el instante $t_{(j)}$, es decir, conocidos los valores marginales n_{A_j}, n_{B_j} y d_j, los cuatro frecuencias observadas en la tabla correspondiente al instante $t_{(j)}$, quedan completamente determinados dada una de ellas, por ejemplo, d_{A_j}. En esas condiciones, si H_o es cierta, d_{A_j} sigue una ley hipergeométrica de parámetros n_j, n_{A_j} y dj, cuya esperanza y varianza son:

$$E[d_{A_j}] = \varepsilon_{A_j} = n_{A_j} \frac{d_j}{n_j}, \tag{2.15}$$

$$Var[d_{A_j}] = \nu_{A_j} = \frac{n_{A_j} n_{B_j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}. \tag{2.16}$$

A patrir de (2.15) obtenemos los valores totales esperados de muertes en los dos grupos (A y B), dados por:

$$E_A = \sum_{j=1}^{J} \varepsilon_{A_j} \quad \text{y} \quad E_B = d - E_A,$$

donde d es el número total de muertes observados considerando los dos grupos.
La estadística de *log-rank* para probar la hipótesis de igualdad entre las dos funciones de supervivencia es dado por

$$U^2 = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B}, \qquad (2.17)$$

donde O_A y O_B representan los números totales observados de muertes en cada grupo. La expresión dada en (2.17) sigue la distribución de Chi-cuadrado (χ_1^2) con un grado de libertad.

Si el resultado es estadísticamente significativo (un p-valor pequeño), concluiremos que los dos grupos no presentan las misma supervivencia. Es decir, la existencia de la desigualdad entre las curvas de supervivencia, nos permitirá determinar una cantidad interesante, llamado *riesgo relativo* y es, estimado por:

$$RR = \frac{O_A / E_A}{O_B / E_B}. \qquad (2.18)$$

La cantidad dada en (2.18) representa la proporcionalidad de un grupo con relación al otro grupo.

Ejemplo 7: Para presentar los resultados anteriores, utilizando el lenguaje R, tomaremos los datos del ejemplo dado en la página 29.

```
> library(survival)

>tiempo<-
c(1,2,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23,6,6,6,6,7,9,10,10,11,13,16,17,19,
20,22,23,
```

25,32,32,34,35)

>estado<c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,0,1,0,0,1,1,0,0,0,1,1,0,0,0,0,0)

Tabla 2.7: Según R se tiene El grupo de Placebo y 6-PM.

	tiempo	estado	grupo
1	1	1	placebo
2	2	1	placebo
3	2	1	placebo
4	2	1	placebo
5	3	1	placebo
6	4	1	placebo
7	4	1	placebo
8	5	1	placebo
9	5	1	placebo
10	8	1	placebo
11	8	1	placebo
12	8	1	placebo
13	8	1	placebo
14	11	1	placebo
15	11	1	placebo
16	12	1	placebo
17	12	1	placebo
18	15	1	placebo
19	17	1	placebo
20	22	1	placebo
21	23	1	placebo
22	6	1	6-MP
23	6	1	6-MP
24	6	1	6-MP
25	6	0	6-MP
26	7	1	6-MP
27	9	0	6-MP
28	10	1	6-MP
29	10	0	6-MP
30	11	0	6-MP
31	13	1	6-MP
32	16	1	6-MP
33	17	0	6-MP
34	19	0	6-MP
35	20	0	6-MP

36	22	1	6-MP
37	23	1	6-MP
38	25	0	6-MP
39	32	0	6-MP
40	32	0	6-MP
41	34	0	6-MP
42	35	0	6-MP

>grupo<c("placebo","6-MP","6-MP")

> leucemia<-data.frame(tiempo,estado,grupo)

> leucemia

> survdiff(Surv(tiempo,estado)~grupo,data=leucemia)

Call:
survdiff(formula = Surv(tiempo, estado) ~ grupo, data = leucemia)

Tabla 8: Obtención de valores observados de 6-MP y Placebo.

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
grupo=6-MP	21	9	19.3	5.46	16.8
grupo=placebo	21	21	10.7	9.77	16.8

Chisq= 16.8 on 1 degrees of freedom, p= 4.2e-05.

En base a los valores obtenidos en la tabla 2.8, tenemos que el valor estadístico

$$U^2 = 16.8 > 5.46+9.77=15.23$$

con un p-valor igual a 4.17e-05. En consecuencia, podemos concluir que existe clara evidencia de desigualdad entre las curvas de supervivencia (Louzada Neto, Mazucheli y Achcar:2002).

La cantidad

$$RR= \frac{9/19.2}{21/10.8} = \frac{0.47}{1.94} =0.24.$$

Esto significa, que la eficacia de la droga 6-mercaptopurina (6-MP) es 0.24 veces que a los pacientes de control de prolongar el estado de remisión de leucemia.

CAPITULO III:

MODELO DE REGRESIÓN DE COX.

3.1.- INTRODUCCIÓN.

Es 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. Se trata por tanto de calcular la tasa de mortalidad como una función del tiempo y de las variables pronósticos. Como en muchas áreas de la estadística, los métodos que se utilizan en análisis de supervivencia son tanto paramétricas como no paramétricas; entendiéndose por estos últimos, aquellos métodos que no suponen ninguna distribución sobre los datos observados. También existen modelos semi-paramétricos, tal como expondremos en este capítulo.

3.2.- MODELOS DE REGRESIÓN DE COX.

El Modelo de Regresión de Cox, propuesto por Cox en 1972, es sin duda uno de los modelos estadísticos más usuales en el análisis de datos de supervivencia. Este modelo permite que el estudio de los tiempos de vida hasta la ocurrencia de un evento de interés se realice considerando variables de interés X denominados covariables. De la misma forma la función de supervivencia en cualquier instante t queda modelada en presencia de covariables.

La expresión general del modelo de regresión de Cox es

$$\lambda(t,X)=\lambda_0(t)Exp(\beta^T X),$$

donde t es el tiempo de vida y X es el vector de covariables.

Supongamos

$$t_i,\ i=1,2,...,k\leq n \qquad (3.1)$$

son independientes (tiempos de falla) y, que el riesgo del j-ésimo individuo es dado por

$$\lambda(t_i/Xj)= \lambda_0(t_i)exp(\beta^T Xj),\ \ j=1,2,...,n, \qquad (3.2)$$

donde $\lambda_0(t_i)$ es conocido como una función de riesgo básico, y β es un vector de dimensión p de coeficientes de regresión no conocido y X_j es el vector de dimensión p de covariables observadas para el j-ésimo individuo.

Observaciones:

a) La función $\lambda_0(t)\geq 0$, representa el riesgo de un individuo con covariables X= **0**.

b) Se conoce como función de riesgo base a $\lambda_0(t)$, porque es una función de riesgo común a todos los individuos y, a la que no se le pone ninguna restricción. Esta es la parte no paramétrica del modelo.

c) La componente

$$exp\{\beta^T X_j\},\ \ j=1,2,..,n,$$

es una función de las covariables del individuo, por lo que tomará un valor distinto para cada individuo. Esta es la parte paramétrica del modelo.

El modelo de Cox, definido en (3.2) es conocido como modelo semi-paramétrica, porque las covariables aparecen como un factor multiplicativo de la función real $\lambda_0(t)\geq 0$, que es, sin restricción alguna. Por otro lado, la definición dada en (3.2) se fundamenta en el hecho de que una razón entre las funciones de riesgo de dos individuos.

$$\frac{\lambda(t/X_i)}{\lambda(t/X_j)}=\frac{\lambda_0(t)\exp\{\beta^T X_i\}}{\lambda_0(t)\exp\{\beta^T B_j\}}=\exp\{\beta^T(X_i-X_j)\}, \qquad (3.3)$$

para i, j =1,2,...,n e $i \neq j$, no depende de t, o sea, razón dada en (3.3) es constante para cada valor t.

Observación: La constante β_0 no aparece en el componente paramétrica $\exp\{\beta^T X_j\}$, debido a que el componente no-paramétrico absorbe este término constante.

(LA PALABRA: EN LA VERSION PDF, **SUPOSICION BASICA** PONER SIN NEGRITA)

La suposición básica para el uso del modelo de riesgos proporcionales de Cox, es que, las tasas de falla sean proporcionales. En otras palabras, que, si un individuo se presenta al inicio de estudio con un riesgo igual 3 veces el riesgo de otro individuo, entonces ésta será la misma para cualquier tiempo t durante el periodo de estudio.

El logaritmo del cociente de las funciones de riesgo se relaciona linealmente con los factores pronósticos (covariables), mientras que en una regresión lineal es la media de la variable respuesta la que se relaciona linealmente con los factores pronósticos.

Nosotros sabemos que la función de supervivencia

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

Entonces

$$S_0(t)=\exp\left(-\int_0^t \lambda_0(s)ds\right),$$

donde

$$\int_0^t \lambda_0(s)ds = \Lambda_0(t) \Rightarrow S_0(t)=\exp(-\Lambda_0(t)), \qquad (3.4)$$

conocido $\Lambda_0(t)$ como función de riesgo base acumulado, de modo que

$$S(t/X)= \exp\left(-\int_0^t \lambda_0(s)ds \ \exp(\beta^T X)\right)ds$$

$$= \exp\left(-\int_0^t \lambda_0(s)ds \ \exp(\beta^T X)\right)$$

$$= \exp(-\Lambda_0(s)\exp(\beta^T X))$$

$$S(t/X)=(S_0(t))^{\exp(\beta^T X)} \qquad (3.5)$$

Este resultado nos permite encontrar la función de densidad f(t/X), puesto que:

$$\lambda(t/X)=\frac{f(t/X)}{S(t/X)} \Rightarrow f(t/X)= \lambda\,(t/X).S(t/X)$$

$$= \lambda_0(t)\exp(\beta^T X)S(t/X)$$

$$= \lambda_0(t)\exp(\beta^T X)\,(S_0(t))^{\exp(\beta^T X)},$$

y la gráfica de $S_0(t)$ comparado con la gráfica de $(S_0(t))^{\exp(\beta^T X)}$ tiene el comportamiento siguiente.

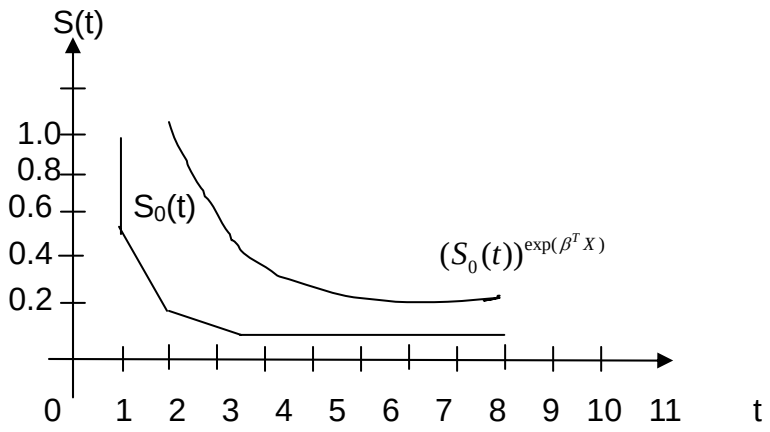

Figura 14: Gráfica comparativa de $S_0(t)$ y S(t).

Esta gráfica nos indica que la suposición de riesgos proporcionales en el modelo de Cox se satisface, porque las curvas son aproximadamente paralelas.

3.2.1.- ESTIMACION DE LOS PARÁMETROS EN EL MODELO DE COX.

El modelo de Cox se caracteriza por los coeficientes β que deben ser estimados a partir de las observaciones muestrales. La presencia del componente no-paramétrica inválida el uso del método de máxima verosimilitud.

Para estimar el vector de parámetros β. Cox (1972) propuso, el método de la función de verosimilitud parcial $L(\beta)$, lo cuál se construye a partir de una

muestra de n individuos, y, ocurran k fallas distintas (k≤n) en los tiempos t_1, t_2,...,t_k. Entonces

P(el individuo i-ésimo falla en t_i a condición de que uno de los individuos vivos de $R(t_i)$ falle en ese momento)

Esto es equivalente a decir que:

$$P(\text{individuo falla en } t_i / \text{del grupo falla en } t_i) = \frac{P(individuo\ falla\ en\ t_i / estaba\ vivo\ antes)}{P(un\ fallo\ en\ t_i\ del\ grupo\ en\ riesgo / estaba\ vivo\ antes)}$$

$$= \frac{\lambda(t_i / X_i)}{\sum_{j \in R(t_i)} \lambda(t_i / X_j)}$$

$$= \frac{\lambda_0(t_i)\exp(\beta^T X_i)}{\sum_{j \in R(t_i)} \lambda_0(t_i)\exp(\beta^T X_j)}$$

$$= \frac{\exp(\beta^T X_i)}{\sum_{j \in R(t_i)} \exp(\beta^T X_j)}, \qquad (3.6)$$

donde $R(t_i)$ es el conjunto de individuos en riesgo antes de t_i .

Ahora si, la función de verosimilitud para β es el producto sobre todas las fallas

t_i , i=1,2,...,n.

Esto es;

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta^T X_i)}{\sum_{j \in R(t_i)} \exp(\beta^T X_j)} \right)^{\delta_i} \qquad (3.7)$$

donde

$$\delta_i = \begin{cases} 1\ si\ t_i\ no\ es\ censurado \\ 0\ si\ t_i\ es\ censurado \end{cases}, \quad i=1,2,...,n,$$

luego, aplicando logaritmo a (3.7) tendremos:

$$\ln[L(\beta)]=\sum_{i=1}^{n}\delta_i \ln\left(\frac{\exp(\beta^T X_i)}{\sum_{j\in R(t_i)}\exp(\beta^T X_j)}\right)$$

$$=\sum_{i=1}^{n}\left[\delta_i\left(\ln(\exp(\beta^T X_i))-\ln(\sum_{j:R(t_i)}\exp(\beta^T X_j))\right)\right]$$

$$\ell(\beta)=\ln[L(\beta)]=\sum_{i=1}^{n}\delta_i\left[\beta^T X_i-\ln\left(\sum_{j:R(t_i)}\exp(\beta^T X_j)\right)\right]. \qquad (3.8)$$

De donde, derivando a (3.8) respecto a los parámetros β^Ts se tiene:

$$\frac{\partial\ell(\beta)}{\partial\beta_k}=\sum_{i=1}^{n}\delta_i\left[X_{ik}-\frac{\sum_{j\in R(t_i)}X_{ik}\exp(\beta^T X_j)}{\sum_{j\in R(t_i)}\exp(\beta^T X_j)}\right], \quad k=1,2,..,p.$$

Haciendo

$$\frac{\partial\ell(\beta)}{\partial\beta_k}=0,$$

obtendremos el siguiente sistema de ecuación scors (ponderaciones):

$$\sum_{i=1}^{n}\delta_i\left[X_{ik}-\frac{\sum_{j\in R(t_i)}X_{ik}\exp(\beta^T X_j)}{\sum_{j\in R(t_i)}\exp(\beta^T X_j)}\right]=0, \quad k=1,2,..,p. \qquad (3.9)$$

A partir de este sistema de ecuaciones podemos obtener los valores estimados de $\hat{\beta}$.

Ejemplo 8.- Consideremos dos muestra:

Sin tratamiento : 7, 9^+, 18

Con tratamiento : 12, 19^+

y, definimos X de la manera siguiente:

$$X=\begin{cases}0, \ sin \ tratamiento\\ 1, \ con \ tratamiento\end{cases}$$

Para poder obtener la función de máximo verosimilitud parcial, primero ordenamos en forma creciente los tiempos de falla de los dos grupos juntos:

$$\underset{y_{(1)}=t_{(1)}}{7} \quad \underset{y_{(2)}=t_{(2)}}{12} \quad \underset{y_{(3)}=t_{(3)}}{18}$$

luego, los conjunto de riesgo $R_{(i)}$, para i=1,2,3 se tiene:

$$R_{(1)}=\{7,\ 9^{+},12,\ 18,\ 19^{+}\}$$
$$R_{(2)}=\{12,\ 18,\ 19^{+}\}$$
$$R_{(3)}=\{18,\ 19^{+}\}$$

Y, la función de verosimilitud parcial es:

$$L(\beta)=\left[\frac{e^{0\beta}}{e^{0\beta}+e^{0\beta}+e^{\beta}+e^{0\beta}+e^{\beta}}\right]\left[\frac{e^{\beta}}{e^{0\beta}+e^{\beta}+e^{\beta}}\right]\left[\frac{e^{0\beta}}{e^{0\beta}+e^{\beta}}\right]$$
$$=\left[\frac{1}{3+2e^{\beta}}\right]\left[\frac{e^{\beta}}{1+2e^{\beta}}\right]\left[\frac{1}{1+e^{\beta}}\right],$$

de donde;

$$\ell(\beta)=\ln[L(\beta)]=-\ln[3+2e^{\beta}]+\beta-\ln[1+2e^{\beta}]-\ln[1+e^{\beta}]$$

de modo que

$$\frac{d\ell(\beta)}{d\beta}=-\frac{2e^{\beta}}{3+2e^{\beta}}+1-\frac{2e^{\beta}}{1+2e^{\beta}}-\frac{e^{\beta}}{1+e^{\beta}}=0,$$

hagamos $t=e^{\beta}$, para obtener:

$$11t+24t^{2}+12t^{3}=3+11t+12t^{2}+4t^{3},$$

finalmente tendremos.

$$8t^{3}+12t^{2}-3=0.$$

Las raíces de ésta ecuación cúbica son:

$$t=\begin{cases}-1.2660\\-0.6736\\0.4397\end{cases}$$

Los dos valores negativos se descartan, ya que $t=e^{\beta}>0$, por tanto:

$$e^{\beta}=0.4397 \Rightarrow \beta= \ln(0.4397)=-0.823= \hat{\beta}.$$

Así logramos que,

$$\lambda(t/X)= \lambda_0(t)\exp(-0.823X).$$

3.2.2.-ESTIMACION DE $\lambda_0(t)$.

Asumamos que $\hat{\lambda}_0(t)$ es constante entre tiempos de supervivencia no censuradas (Mei-Cheng, 2005). En primer lugar, ordenamos los tiempos de supervivencia no censuradas en forma creciente, así

$$y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(k)}$$

y, el conjunto de riesgos $R_{(j)}$ asociado a estos tiempos ordenados está dada por:

$$R_{(j)}=\{T_i : T_i \geq y_{(j)}\},\ j=1,2,\ldots,k.$$

Sea $\hat{\lambda}_{(0)}$, $\hat{\lambda}_{(1)}$, $\hat{\lambda}_{(2)}$,... constantes, es decir:

$$\hat{\lambda}_0(t)=\begin{cases} \hat{\lambda}_{(0)}\ , 0 \leq t < y_{(1)} \\ \hat{\lambda}_{(1)}\ , y_{(1)} \leq t < y_{(2)} \\ \quad \ldots\ldots\ldots \end{cases}$$

Digamos que queremos estimar $\hat{\lambda}_{(0)}$, el individuo del conjunto de riesgos en $y_{(2)}$ está en el grupo $R_{(2)}$. Como sabemos que una persona falla en $y_{(2)}$ dado ($y_{(2)}$, $R_{(2)}$). Entonces

$$1= \sum_{j \in R_{(2)}} P[el\ j-esimo\ individuo\ falla\ en\ y_{(2)} / y_{(2)}, R_{(2)}]$$

$$= \sum_{j \in R_{(2)}} (y_{(3)} - y_{(2)}) \hat{\lambda}_{(2)} e^{\beta^T X_j}$$, puesto que $\lambda(t)\Delta(t) \simeq P[t \leq T < t+\Delta t / T \geq t]$

$$= (y_{(3)}-y_{(2)}) \hat{\lambda}_{(2)} \sum_{j \in R_{(2)}} e^{\beta^T X_j} .$$

Así, el riesgo de probabilidad entre $y_{(2)}$ y $y_{(3)}$ es

$$(y_{(3)}-y_{(2)}) \hat{\lambda}_{(2)} = \frac{1}{\sum_{j \in R_{(2)}} e^{\beta^T X_j}}, \qquad (3.10)$$

Ahora usamos $\hat{\beta}$ (obtenido mediante el método de máximo verosimilitud parcial) para obtener

$$\hat{\lambda}_{(2)} = \frac{1}{(y_{(3)} - y_{(2)}) \sum_{j \in R_{(2)}} e^{\hat{\beta}^T X_j}} = \hat{\lambda}_0(t), \qquad (3.11)$$

con lo cual queda estimado el riesgo base de $\lambda_0(t)$. Por otro lado, este resultado nos permite obtener también, la estimación de función de riesgo acumulado $\hat{\Lambda}_0(t)$.

3.3.- INTERPRETACIÓN DE LOS PARÁMETROS EN EL MODELO DE COX.

Los coeficientes β en el modelo de regresión de Cox, miden los efectos de la covariables sobre la tasa de falla; siendo que , una covariable puede acelerar, o desacelerar la función de riesgo. Puesto que, el modelo de riesgo proporcional general para el i-ésimo individuo es

$$\lambda_i(t/X_i) = \lambda_0(t) \exp\left(\sum_{j=1}^{p} \beta_j X_{ij}\right), \; i=1,2,...,n, \qquad (3.12)$$

entonces

$$\frac{d \ln \lambda_i(t / X_i)}{dX_{ij}} = \beta j \ , \ j=1,2,...,p \ ; \ i=1,2,...,n$$

o sea, β_j da el cambio proporcional en la función de riesgo que resulta de un cambio marginal en la j-ésima variable explicativa.

Obsérvese que la tasa de falla para dos individuos (i y l) que presentan los mismos valores de covariables (covariables fijos), excepto en la j-ésima de ellas, es dada por

$$\frac{\lambda(t / X_i)}{h(t / X_l)} = \frac{\exp(\beta_j^T X_{ij})}{\exp(\beta_j^T X_{lj})} = \exp\{\beta_j(X_{ij}-X_{lj})\}.$$

Esta razón, se conoce como **riesgo relativo,** es constante en el tiempo y las tasas de riesgo son proporcionales, si la j-ésima covariables fuese dicotómica (variable binaria) en que $X_{ij}=1$ y $X_{lj}=0$, entonces

$$\frac{\lambda(t / X_i)}{\lambda(t / X_l)} = \exp\{\beta_j\}.$$

Esto es, el riesgo de falla del individuo i-ésimo es $\exp\{\beta_j\}$ veces el riesgo de falla de individuo j-ésimo; para el caso en que las demás covariables se mantienen fijas (en particular, cuando las variables son dicotómicas).

Ejemplo 9.- Feigl y Zelen (Abaurrea y Cebrián-2004), analizaron la supervivencia de dos grupos de enfermos de leucemia clasificados, en base a una característica celular, en ag positivos (ag^+) y ag negativos (ag^-), tabla 3.1. Los tiempos de fallo corresponden al tiempo, medido en semanas, desde el instante del diagnóstico hasta el fallecimiento; también se registró la cantidad de glóbulos blancos contagiados (o contaminados) de cada paciente en el momento del diagnóstico.

Tabla 9: Valores de ag positivos y ag negativos.

ag positivos, N=17			**ag negativos, N=16**		
N° Glob. B (wbc)	**log(wbc)**	**T. Superv.**	**N° Glob. B (wbc)**	**log(wbc)**	**T. Superv.**
2 300	3.36	65	4 400	3.64	56
750	2.88	156	3 000	3.48	65

4 300	3.63	100	4 000	3.60	17
2 600	3.41	134	1 500	3.18	7
6 000	3.78	16	9 000	3.95	16
10 500	4.02	108	5 300	3.72	22
10 000	4.00	121	10 000	4.00	3
17 000	4.23	4	19 000	4.28	4
5 400	3.73	39	27 000	4.43	2
7 000	3.85	143	28 000	4.45	3
9 400	3.97	56	31 000	4.49	8
32 000	4.51	26	26 000	4.41	4
35 000	4.54	22	21 000	4.32	3
100 000	5.00	1	79 000	4.90	30
100 000	5.00	1	100 000	5.00	4
52 000	4.72	5	100 000	5.00	43
100 000	5.00	65			

Aplicaremos el software R para estimar los coeficientes de Regresión de Cox. Así:

```
> library(survival)
> tiemp1<-c(65,156,100,134,16,108,121,4,39,143,56,26,22,1,1,5,65)
> cens1<-rep(1,17)
>lwbc1<c(3.36,2.88,3.63,3.41,3.78,4.02,4.00,4.23,3.73,3.85,3.97,4.51,4.54,5.0
0,5.00,4.72,5.00)
> tiemp2<-c(56,65,17,7,16,22,3,4,2,3,8,4,3,30,4,43)
> cens2<-rep(1,16)
>lwbc2<c(3.64,3.48,3.6,3.18,3.95,3.72,4.0,4.28,4.43,4.45,4.49,4.41,4.32,4.90,5
.0,5.0)
> tiemp<-c(tiemp1,tiemp2)
> cens<-c(cens1,cens2)
> lwbc<-c(lwbc1,lwbc2)
> ag<-c(rep(0,17),rep(1,16))
> leuc<-as.data.frame(cbind(tiemp,cens,ag,lwbc))
> leuc
```

Tabla 10: Datos observados según el lenguaje **R**.

	tiemp	cens	ag	lwbc
1	65	1	0	3.36
2	156	1	0	2.88
3	100	1	0	3.63
4	134	1	0	3.41
5	16	1	0	3.78
6	108	1	0	4.02
7	121	1	0	4.00
8	4	1	0	4.23
9	39	1	0	3.73
10	143	1	0	3.85
11	56	1	0	3.97
12	26	1	0	4.51
13	22	1	0	4.54
14	1	1	0	5.00
15	1	1	0	5.00
16	5	1	0	4.72
17	65	1	0	5.00
18	56	1	1	3.64
19	65	1	1	3.40
20	17	1	1	3.60
21	7	1	1	3.18
22	16	1	1	3.95
23	22	1	1	3.72
24	3	1	1	4.00
25	4	1	1	4.28
26	2	1	1	4.43
27	3	1	1	4.45
28	8	1	1	4.49
29	4	1	1	4.41
30	3	1	1	4.32
31	30	1	1	4.90
32	4	1	1	5.00
33	43	1	1	5.00

> ajuste<-coxph(Surv(tiemp,cens)~ag+lwbc,data=leuc)

> summary(ajuste)

Tabla 11: Coeficientes de Cox obtenido según el lenguaje **R**.

Call: coxph(formula = Surv(tiemp, cens) ~ ag + lwbc, data = leuc)

n= 33

	coef	exp(coef)	se(coef)	z	p
ag	1.069	2.91	0.429	2.49	0.013
lwbc	0.844	2.33	0.313	2.70	0.007

Por consiguiente:

$$\lambda(t/X_i)= \lambda_0(t)\exp[1.069(ag_{i1}) + 0.844(lwbc_{i2})],\ i=1,2,\ldots, 33,$$

de donde

$$\ln\left[\frac{\lambda(t/X_i)}{\lambda_0(t)}\right] = 1.069(ag_{i1}) + 0.844(lwbc_{i2}),\ \ i=1,2,\ldots, 33.$$

Y

$$S(t/X_i)= [S_0(t)]^{\exp\{1.069(ag_{i1})+0.844(lwbc_{i2})\}}$$

Además en la tabla 11 tenemos también los riesgos relativos (a partir de los **Exp(coef)),** con los cuales podemos decir que el efecto de los ag tiene un riesgo relativo de morir 2.91 veces el de positivo respecto al de negativo. En cuanto al logaritmo del contagio al número de glóbulos blancos (lwbc), una persona con una cierta cantidad de glóbulos blancos contagiados tiene 2.33 veces el riesgo de morir en relación a una persona con una unidad menor de contagiado. La columna *se(coef)* representa el error estándar de las respectivas covariables con su respectivo z valor normalizada que para el caso de ag vale 2.49 y para lwbc es 2.70. Finalmente en esta tabla tenemos los p-valores respectivos, que para ambos covariables resultan menores de 0.05, por lo que se rechaza la hipótesis de que los coeficientes de ambas covariables sean iguales a cero.

3.4.- SUPOSICIÓN DE RIESGOS PROPORCIONALES EN EL MODELO DE

COX.

El modelo de Cox es adecuado para situaciones en que la suposición de riesgos proporcionales es válida; en otras palabras, esto es lo mismo decir; la *verificación de la proporcionalidad en los riesgos*, y la metodología consiste en situaciones en que las funciones de riesgo no se interceptan. Este hecho se puede verificar gráficamente de la manera siguiente:

i) Dividir los datos en j estratos de acuerdo con las j categorías de alguna variable. Por ejemplo, en dos estratos (j=2) si la covariable considerado es sexo.

ii) Estimar $\hat{\Lambda}_{0j}(t)$ para cada estrato j obteniéndose las curvas

$$\log[\hat{\Lambda}_{0j}(t)]=\log[-\log(S(t))] \qquad (3.13)$$

versus t, o log(t).

Si la suposición es válida, las curvas de $\log[\hat{\Lambda}_{0j}(t)]$ versus t, o log(t), deben presentar diferencias constantes en el tiempo, o sea, deben ser aproximadamente paralelas. Estas curvas pueden ser obtenidas para cada covariable en estudio, de este modo, encontrar indicios cual covariable no satisface la suposición de riesgos proporcionales (Victor Moreno-2004).

El hecho de que la expresión dada en (3.13) representa aproximadamente curvas paralelas, formalmente se debería justificar así. Por (3.2) sabemos que

$$\lambda(t/Xj)= \lambda_0(t)\exp(\beta^T Xj),\ \ j=1,2,...,n,$$

entonces

$$\int_0^t \lambda_i(s/X_i)ds = \int_0^t \lambda_0(s)\exp(\beta^T X_i)ds\,.$$

Luego

$$\Lambda(t_j) = \Lambda_0(t_i)\exp(\beta^T X_j)=e_j\ ,\ j=1,2,...,n \qquad (3.14)$$

y,

$$\ln(\Lambda(t_j))=\ln(\Lambda_0(t_i))+\beta^T X_j\ ,\ \ j=1,2,...,n. \qquad (3.15)$$

Por otro lado sabemos que:

$$\Lambda(t)=-\ln(S(t)),$$

Entonces en (3.15) se tiene:

$$\ln(-\ln(S(t_j)))= \ln(-\ln(S_0(t_j))) +\beta^T X_j \ , \ j=1,2,...,n \qquad (3.16)$$

Luego, según la fórmula (3.16) las curvas en cada grupo seguirán la forma de la supervivencia base S_0 y aproximadamente se mantendrán paralelas, separadas por la distancia marcada por el coeficiente β.Tal como se muestra en la figura 15.

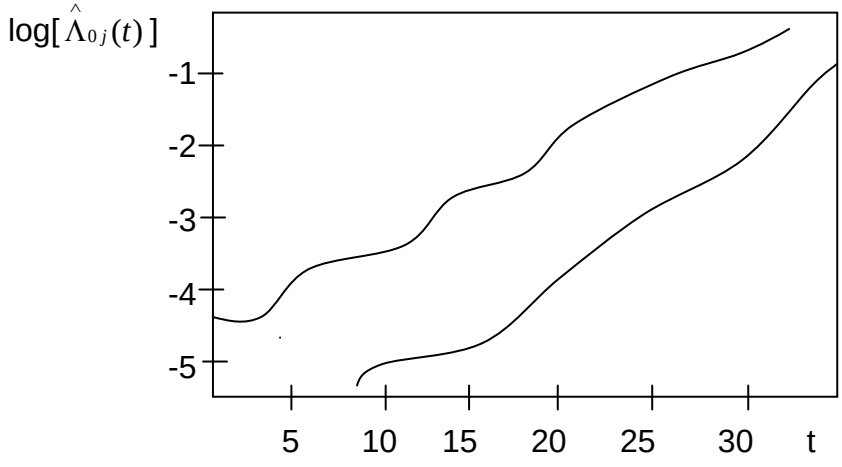

Figura 15: Comparación de las curvas riesgos log[$\hat{\Lambda}_{0j}(t)$] vs t en días.

3.5.- LOS RESIDUOS EN EL MODELO DE COX.

En regresión lineal es usual verificar la adecuación del modelo ajustado por medio de la inspección de gráficos de los residuos. En análisis de supervivencia, debido a la presencia de datos de censura, los residuales no seguirán la distribución normal y pueden ser altamente asimétricos. Sin

embargo, existen otros métodos alternativos de residuos para validar el modelo.

3.5.1.-RESIDUOS DE COX-SNELL.

Para los n individuos en estudio, los residuos de Cox-Snell (1968), para el modelo de Cox, se define por:

$$e_i = \hat{\Lambda}_0(t_i)\exp(\sum_{k=1}^{p} x_{ik}\,\hat{\beta}_k)\text{, } i=1,2,...,n, \qquad (3.17)$$

y se usa para verificar la calidad de ajuste del modelo de Cox. Si el modelo esta bién ajustado, el gráfico $\hat{\Lambda}(e_i)$ versus e_i es aproximadamente una recta. Los rediduos de Cox-Snell son usados para examinar el ajuste global del modelo de Cox.

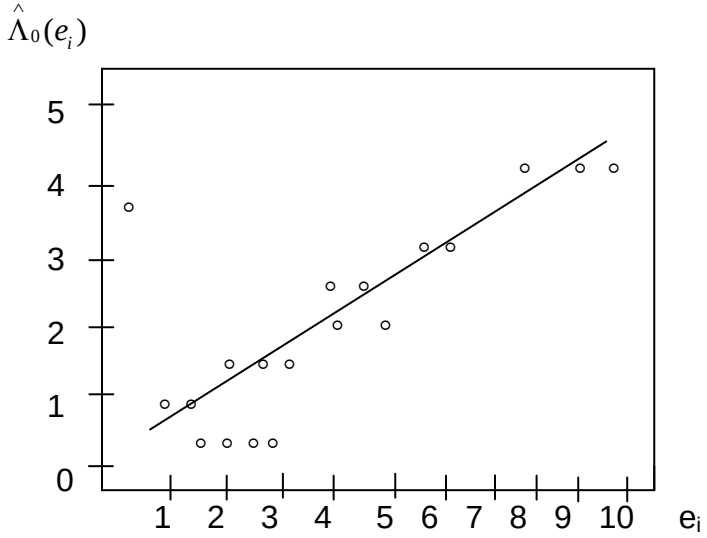

Figura 16: Gráfica de Cox-Snell de $\hat{\Lambda}(e_i)$ versus e_i

3.5.2.- RESIDUOS DE SCHOENFELD.

Consideremos los k ($k \leq n$) tiempos diferentes de falla t_1, t_2 ,...,t_k . Un vector de residuos de schoenfeld es obtenido en cada tiempo observado de falla. Si el individuo i-ésimo es observado fallar, el correspondiente residuo se define por:

$$r_i = X_i - E[\text{covariable en } t_i \,/\, R_{(t_i)}]$$

$$= X_i - \frac{\sum_{j \in R_{(t_i)}} X_j e^{\hat{\beta}^T X_j}}{\sum_{j \in R_{(t_i)}} e^{\hat{\beta}^T X_j}}, \quad i=1,2...,k. \tag{3.18}$$

En (3.18) se propone que:

$$E[X_i/R_{(i)}] = \frac{\sum_{j \in R_{(t_i)}} X_j e^{\hat{\beta}^T X_j}}{\sum_{j \in R_{(t_i)}} e^{\hat{\beta}^T X_j}}, \quad i=1,2,...,k \tag{3.19}$$

ya que, poniendo $p(X_i)$ como

$$p(X_i) = \frac{e^{\beta^T X_j}}{\sum_{j \in R_{(i)}} e^{\beta^T X_j}} > 0, \quad i=1,2,...,k$$

y,

$$\sum_{j \in R_{(i)}} p(X_j) = \sum_{j \in R_{(i)}} \left[\frac{e^{\beta^T X_j}}{\sum_{j \in R_{(i)}} e^{\beta^T X_j}} \right] = 1$$

Se tiene que, $p(X_i)$ satisface las condiciones de una función de probabilidad.

Como tenemos en (3.15) el valor esperado de la covariable X_i dado $R_{(t_i)}$, en el momento de fallo en t_i, de ese modo la interpretación de r_i dado en (3.15) tiene sentido. Esos residuos, dibujados frente al tiempo observado de supervivencia deben repartirse aleatoriamente alrededor de 0 ($\sum_i r_i \simeq 0$), si el modelo de riesgos proporcionales es correcto. Si la gráfica mostrara alguna tendencia, habría que incluir en el modelo alguna covariable dependiente del tiempo.

Observaciones:

1.- El residual r_i es el vector con un componente de cada covariable, condicionada una única falla en el conjunto de riesgo $R_{(t_i)}$.

2.- El residual r_i para i=1,2,...,k; representa la contribución individual de la derivada del verosimilitud parcial logarítmica.

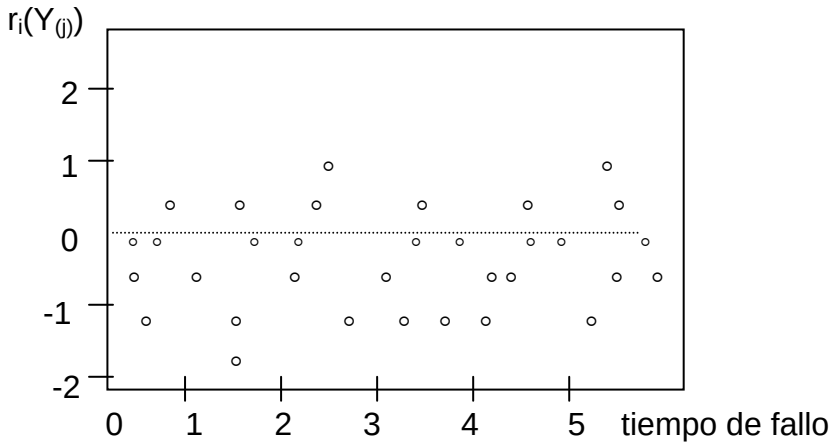

Figura 17: Representación gráfica de residuos de Schoenfeld

El valor esperado de la falla de un individuo en estudio es expresado por el término:

$$\frac{\sum_{j \in R_{(t_i)}} X_j e^{\hat{\beta} X_j}}{\sum_{j \in R_{(t_i)}} e^{\hat{\beta} X_j}}, \quad i=1,2...,k$$

y, en (3.18) para i=1,2,...,k tendremos

$$r_i = \frac{X_i \sum_{j \in R_{(t_i)}} e^{\hat{\beta} X_j} - \sum_{j \in R_{(t_i)}} X_j e^{\hat{\beta} X_j}}{\sum_{j \in R_{(t_i)}} e^{\hat{\beta} X_j}},$$

entonces

$$\sum_{i=1}^{k} r_i = \frac{\sum_{i=1}^{k} X_i \sum_{j \in R_{(t_i)}} e^{\hat{\beta} X_j} - \sum_{i=1}^{k} \sum_{j \in R_{(t_i)}} X_j e^{\hat{\beta} X_j}}{\sum_{j \in R_{(t_i)}} e^{\hat{\beta} X_j}} \simeq 0.$$

3.5.3.- RESIDUOS DE MARTINGALE.

Los residuos de Cox-Snell, coinciden con los valores de la función de riesgo acumulada. Es decir; si

$$\lambda(t/X_i) = \lambda_0(t)\exp(\beta^T X_i),\ i=1,2,...,n$$

entonces

$$\int_0^t \lambda_i(s / X_i)ds = \int_0^t \lambda_0(s)\exp(\beta^T X_i)ds\,.$$

Luego

$$\Lambda(t_i) = \Lambda_0(t_i)\exp(\beta^T X_i) = e_i\ ,\ i=1,2,...,n. \qquad (3.20)$$

Para $t=t_i$ (tiempo de falla). La expresión dada en (3.20) se llama **función de impacto.** Sin embargo, resulta mucho más útil una sencilla transformación de los mismos: uno menos la función de impacto, si el dato es no censurado; menos la función de impacto si el dato es censurado. Esto es:

$$M_i = \delta_i - \hat{\Lambda}_0(t_i)\exp(\sum_{k=1}^{p} x_{ik}\beta_k) = \delta_i - e_i,\ \ i=1,2,...,n.$$

Los valores así obtenidos se le llama **residuos de Martingales.**

La gráfica de estos residuos frente a alguna covariable se interpreta como en regresión múltiple: Una tendencia en la gráfica indica la necesidad de incorporar alguna transformación de la covariable en el modelo. También sirve para detectar posibles observaciones aberrantes (outliers).

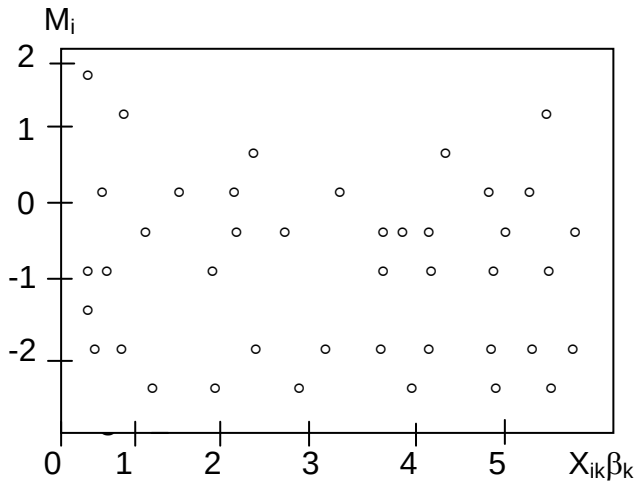

Figura 18: Representación gráfica de Residuos de Martingale

Observación: Estos residuales, generalmente, se aplican cuando los covariables son continuos.

Sin embargo, cuando las covariables son cualitativos, la forma de verificar gráficamente la hipótesis de riesgos proporcionales en el modelo de Cox, consiste en representar ln(-ln S(t)) en función de ln(t) para cada una de las categorías (o estratos). Si se cumple la hipótesis de riesgos proporcionales, éstas gráficas tienen que ser aproximadamente paralelas.

4.6.- APLICACIÓN.

En la Tabla 12, presentamos información de 65 pacientes sobre tiempo t de supervivencia dada en meses y, las respectivas medidas en 5 covariables de cada paciente aplicados al momento del diágnosis (Lawless -1982).

- X_1 : Logaritmo de urea en sangre.
- X_2 : Hemoglobina.
- X_3 : Edad.
- X_4 : Sexo (0=hombres, 1= mujeres).
- X_5 : Calcio.

Tabla 12: Datos observados de 65 pacientes.

N^o	t	X_1	X_2	X_3	X_4	X_5	N^o	t	X_1	X_2	X_3	X_4	X_5
1	1	2.218	9.4	67	0	10	34	26	1.230	11.2	49	1	11
2	1	1.940	12.0	38	0	18	35	32	1.322	10.6	46	0	9
3	2	1.519	9.8	81	0	15	36	35	1.114	7.0	48	0	10
4	2	1.748	11.3	75	0	12	37	37	1.602	11.0	63	0	9
5	2	1.301	5.1	57	0	9	38	41	1.000	10.2	69	0	10
6	3	1.544	6.7	46	1	10	39	42	1.146	5.0	70	1	9
7	5	2.236	10.1	50	1	9	40	51	1.568	7.7	74	0	13
8	5	1.681	6.5	74	0	9	41	52	1.000	10.1	60	1	10
9	6	1.362	9.0	77	0	8	42	54	1.255	9.0	49	0	10
10	6	2.114	10.2	70	1	8	43	58	1.204	12.1	42	1	10
11	6	1.114	9.7	60	0	10	44	66	1.447	6.6	59	0	9
12	6	1.415	10.4	67	1	8	45	67	1.322	12.8	52	0	10
13	7	1.978	9.5	48	0	10	46	88	1.176	10.6	47	1	9
14	7	1.041	5.1	61	1	10	47	89	1.322	14.0	63	0	9
15	7	1.176	11.4	53	1	13	48	92	1.431	11.0	58	1	11
16	9	1.724	8.2	55	0	12	49	4^+	1.945	10.2	59	0	10
17	11	1.114	14.0	61	0	10	50	4^+	1.924	10.0	49	1	13
18	11	1.230	12.0	43	0	9	51	7^+	1.114	12.4	48	1	10
19	11	1.30	13.2	65	0	10	52	7^+	1.532	10.2	81	0	11
20	11	1.508	7.5	70	0	12	53	8^+	1.079	9.9	57	1	8
21	11	1.079	9.6	51	1	9	54	12^+	1.146	11.6	46	1	7
22	13	0.778	5.5	60	1	10	55	11^+	1.613	14.0	60	0	9
23	14	1.398	14.6	66	0	10	56	12^+	1.398	8.8	66	1	9
24	15	1.602	10.6	70	0	11	57	13^+	1.663	4.9	71	1	9
25	16	1.342	9.0	48	0	10	58	16^+	1.146	13.0	55	0	9
26	16	1.322	9.8	62	1	10	59	19^+	1.322	13.0	59	1	10
27	17	1.230	10.0	53	0	9	60	19^+	1.322	10.8	69	1	10
28	17	1.591	11.2	68	0	10	61	28^+	1.230	7.3	82	1	9
29	18	1.447	7.5	65	1	8	62	41^+	1.756	12.8	72	0	9
30	19	1.079	14.4	51	0	15	63	53^+	1.114	12.0	66	0	11
31	19	1.255	7.5	60	1	9	64	57^+	1.255	12.5	66	0	11
32	24	1.301	14.6	56	1	9	65	77^+	1.079	14.0	60	0	12
33	25	1.000	12.4	67	0	10							

> library(survival)
>tiempo<c(1,1,2,2,2,3,5,5,6,6,6,6,7,7,7,9,11,11,11,11,11,13,14,15,16,16,17,17,
18,19,19,24,25,26,32,35,37,41,42,51,52,54,58,66,67,88,89,92,4,4,7,7,8,12,11,1
2,13,16,19,19,28,41,53,57,77)

> cens<-c(rep(1,48),rep(0,17))

>urea<c(2.218,1.940,1.519,1.748,1.301,1.544,2.236,1.681,1.362,2.114,1.114,1.415,1.978,1.041,1.176,1.724,1.114,1.230,1.301,1.508,1.079,0.778,1.398,1.602,1.342,1.322,1.230,1.591,1.447,1.079,1.255,1.301,1.000,1.230,1.322,1.114,1.602,1.000,1.146,1.568,1.000,1.255,1.204,1.447,1.322,1.176,1.322,1.431,1.945,1.924,1.114,1.532,1.079,1.146,1.613,1.398,1.663,1.146,1.322,1.322,1.230,1.756,1.114,1.255,1.079)

>hemog<c(9.4,12.0,9.8,11.3,5.1,6.7,10.1,6.5,9.0,10.2,9.7,10.4,9.5,5.1,11.4,8.2,14.0,12.0,13.2,7.5,9.6,5.5,14.6,10.6,9.0,8.8,10.0,11.2,7.5,14.4,7.5,14.6,12.4,11.2,10.6,7.0,11.0,10.2,5.0,7.7,10.1,9.0,12.1,6.6,12.8,10.6,14.0,11.0,10.2,10.0,12.4,10.2,9.9,11.6,14.0,8.8,4.9,13.0,13.0,10.8,7.3,12.8,12.0,12.5,14.0)

>edad<c(67,38,81,75,57,46,50,74,77,70,60,67,48,61,53,55,61,43,65,70,51,60,66,70,48,62,53,68,65,51,60,56,67,49,46,48,63,69,70,74,60,49,42,59,52,47,63,58,59,49,48,81,57,46,60,66,71,55,59,69,82,72,66,66,60)

>sexo<c(0,0,0,0,0,1,1,0,0,1,0,1,0,1,1,0,0,0,0,0,1,1,0,0,0,1,0,0,1,0,1,1,0,1,0,0,0,0,1,0,1,0,1,0,0,1,0,1,0,1,1,0,1,1,0,1,1,0,1,1,1,0,0,0,0)

>calcio<c(10,18,15,12,9,10,9,9,8,8,10,8,10,10,13,12,10,9,10,12,9,10,10,11,10,10,9,10,8,15,9,9,10,11,9,10,9,10,9,13,10,10,10,9,10,9,9,11,10,13,10,11,8,7,9,9,9,9,10,10,9,9,11,11,12)

> mm<-as.data.frame(cbind(t,urea,hemog,edad,sexo, calcio))
> mm

>ajuste.cox<coxph(Surv(tiempo,cens)~urea+hemog+edad+sexo+calcio,da

ta=mm)

> summary(ajuste.cox)

(ESTIMACIÓN Y VALORACIÓN GLOBAL DEL MODELO DE COX)

Tabla 13: Coeficientes de Cox

Call:

coxph(formula = Surv(tiempo, cens) ~ urea + hemog + edad + sexo +

calcio, data = mm)

n= 65

	coef	exp(coef)	se(coef)	z	p
X_1= urea	1.8146	6.139	0.6528	2.780	0.0054
X_2= hemog	-0.1483	0.862	0.0619	-2.395	0.0170
X_3= edad	-0.0221	0.978	0.0162	-1.358	0.1700
X_4= sexo	-0.1792	0.836	0.3238	-0.553	0.5800
X_5= calcio	0.1431	1.154	0.1020	1.403	0.1600

	exp(coef)	exp(-coef)	lower .95	upper .95
X_1= urea	6.139	0.163	1.708	22.066
X_2= hemog	0.862	1.160	0.764	0.973
X_3= edad	0.978	1.022	0.948	1.010
X_4= sexo	0.836	1.196	0.443	1.577
X_5= calcio	1.154	0.867	0.945	1.409

Rsquare= 0.218 (max possible= 0.991)

Likelihood ratio test =	16	on 5 df,	p=0.00683
Wald test	= 16.9	on 5 df,	p=0.00474
Score (logrank) test =	17.9	on 5 df,	p=0.00305

Tabla de los tres criterios

Por tanto, la función de supervivencia estimada es,

$$\hat{S}(t/X) = \left[\hat{S}_0(t)\right]^{\exp(1.8146X_1 - 0.1483X_2 - 0.0221X_3 - 0.1792X_4 + 0.1431X_5)} \quad (3.21)$$

y,

$$\frac{\hat{\lambda}(t/X)}{\hat{\lambda}_0(t)} = \exp(1.8146X_1 - 0.1483X_2 - 0.0221X_3 - 0.1792X_4 + 0.1431X_5). \quad (3.22)$$

Luego, realizamos las siguientes interpretaciones:

- La columna de *Exp(coef)* representa los riesgos relativos cuando X_j, aumenta una unidad, manteniéndose constante las demás en la comparación de dos individuos con las mismas covariables, por ejemplo; el riesgo relativo de *urea* es 6.139.

- La columna *z* es la división del coeficiente estimado entre su error estándar correspondiente, donde los p-valores (significativos) indican que la covariables X_1= urea y X_2= hemoglobina dan mayor contribución al modelo estimado, mientras que otros covariables no son de mayor preponderancia en el modelo. Al menos la X_4= sexo no contribuye significativamente al modelo, esto decimos así, ya que, su p-valor=0.5800 es muy alto.

- El R^2=0.218 nos indica que, aproximadamente el 21.8% de la variación de los tiempos de supervivencia de individuos con mieloma múltiple t, quedan estadísticamente explicados por las cinco covariables mencionados, si el modelo utilizado fuese como modelo lineal múltiple.

- El p-valor 0.00683<0.01 correspondiente a la razón de verosimilitud (Victor Moreno - 2004), o sea, al Likelihood ratio test nos indica que el ajuste dado en (3.22) sería el modelo lo más adecuado posible considerando globalmente, es decir, considerando todas las covariables pre-determinadas.

- El estadístico de Wald según p=0.00474 nos prueba que cada coeficientes β's del modelo son diferentes de cero, esto quiere decir, que la hipótesis nula,

$$H_0 : \beta_j = 0,\ j=1,2,3,4,5$$

debe ser rechazada.

- El hecho de que la prueba de Score(logrank) =17.9 con p-valor=0.00305<0.01, nos hace ver que existen diferencias en la comparación de curvas de supervivencia de dos o más grupos de individuos. Estos es, si es que agrupamos por alguna razón o estrato, supondremos que tales grupos no son iguales.

(PRUEBA DE PROPORCIONALIDAD EN LOS RIESGOS)

> residuals(ajuste.cox,type="martingale")

> cox.zph(ajuste.cox)

Tabla 14: Los p-valores obtenidos para los datos de la tabla 12

	rho	chisq	p
urea	-0.2431	4.3869	0.0362
hemog	-0.0286	0.0392	0.8431
edad	-0.0379	0.0830	0.7733
sexo	-0.0837	0.3370	0.5616
calcio	-0.1666	2.0799	0.1493
GLOBAL	NA	7.4522	0.1891

> op<-par(mfrow=c(2,4))

> plot(cox.zph(ajuste.cox))

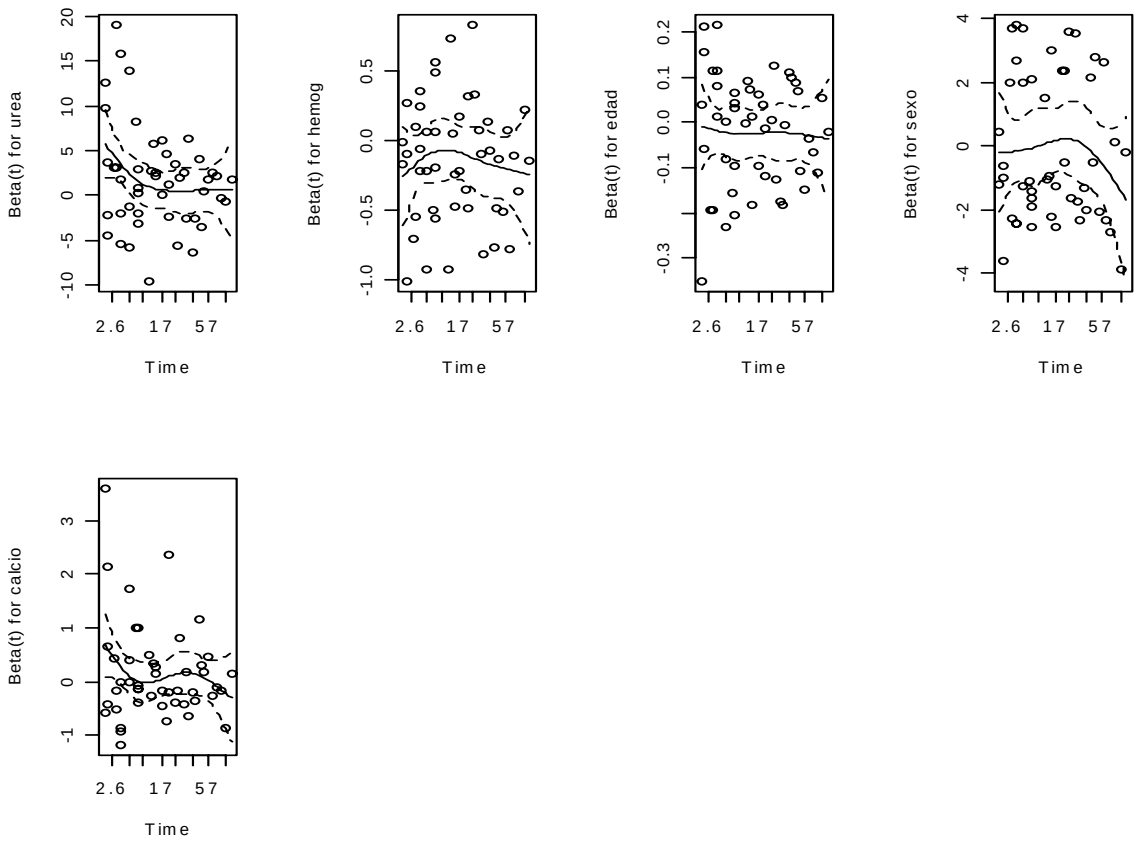

Figura 19: Residuos de Schoenfeld del Modelo de Cox sin interacción.

Las gráficas de la Figura 19, residuos de Schoenfeld para cada individuo y para cada covariables considerada; en los cuales no se observa ningún patrón extraño que revela el rechazo de la suposición de riesgos proporcionales. Es decir en las 5 gráficas anteriores, la presentación de las puntuaciones son totalmente aleatorias alrededor del 0. De la misma forma según los resultados de la tabla 14 podemos decir que no existe evidencia significativa al 5% de que se viole el supuesto de riesgos proporcionales, ni desde el punto de vista global ni para cada covariable.

Por otro lado, el p-valor= 0.00683 (Likelihood ratio test) significa que el sistema de ecuaciones según el método de máxima verosimilitud son solubles, o sea, el sistema esta bien planteadas y tiene solución. El test de Wald nos indica según

p-valor=00474 que los valores estimados de los coeficientes β^s son todos diferentes de cero. Y el Score (logrank) test con p-valor=0.00305 nos indica que las curvas de supervivencia para dos grupos de individuos son diferentes. De los resultados presentados, podemos concluir que la suposición de riesgos proporcionales no se rechaza, ya que, en los gráficos de arriba las tendencias a lo largo del tiempo no son evidentes.

CAPITULO IV:

FORMA FUNCIONAL DE LOS COVARIABLES EN MODELO DE COX

4.1 INTRODUCCION.

Hasta ahora sólo hemos considerado covariables cuyo valor, para cada caso (o sea, para cada individuo), se observa al principio del estudio y permanece constante a lo largo del mismo. Sin embargo, en ocasiones hay que considerar covariables que varían durante el periodo de estudio. Algunos ejemplos son: Ciertas pruebas de laboratorio, urea por ejemplo, pueden repetirse en cada visita del pacientes al nosocomio, así podemos tener observaciones u0, u30, u60, u90,… que podrían corresponder el valor de urea en el momento inicial y, luego, cada treinta días; entonces, la urea sería una covariable dependiente del tiempo. A mitad del seguimiento puede considerarse adecuado cambiar el tratamiento de algún paciente, o sea que, algunos de estos sujetos o individuos cambiasen de grupo de tratamiento. Ésta situación daría lugar a un estrato dependiente del tiempo. Incorporando esas características en el análisis estadístico, obtendremos resultados más precisos en comparación de aquellos resultados que hacen uso solamente de las mismas medidas registradas generalmente al inicio del estudio. Con este tipo de covariables, el Modelo de Cox es extendido para incorporar informaciones de tipo longitudinal.

En este caso, la suposición de riesgos proporcionales es violada y el Modelo de Cox no es adecuado, para lo cual existen modelos alternativos para enfrentar estas situaciones. Uno de ellos es una extensión del propio modelo de Cox, denominado modelo de riesgos proporcionales estratificado y el modelo incluyendo covariables dependientes del tiempo.

4.2 FORMA FUNCIONAL DE LOS COVARIABLES.

Una forma funcional de los covariables, matemáticamente significa que sean funciones que dependan de alguna otra variable, por ejemplo: una covariable dependa del tiempo. En la estructura de Cox, de acuerdo a la naturaleza del problema en estudio, se puede presentarse covariables que tomen una estructura de función diferenciable y continua. Precisamente, este hecho, de manera muy natural es posible adecuar a las covariables en un modelo de Cox. Tema que será abordado en la siguiente sección.

4.3 RIESGO NO PROPORCIONAL DEL MODELO DE COX.

En muchos problemas reales, se hace necesario monitorear las covariables durante el periodo de estudio, ya que, puede haber casos en que los individuos en estudio les conviene un posible cambio de grupo o tratamiento, o en todo caso, la dosis aplicada a algún paciente le propicia un cambio en su estructura relacionado a su supervivencia, y por ende urge un cambio de grupo o un cambio en su tratamiento, lo cual, este hecho debe ser tomada como una covariable cambiante en el tiempo, a fin de obtener resultados más exactos y concretos respecto al análisis estadístico.

4.3.1.- MODELO DE COX CON COVARIABLES DEPENDIENTES DEL TIEMPO.

Las covariables en el modelo de Cox considerados en el capítulo anterior fueron medidos al inicio del estudio o sino al origen del tiempo. Sin embargo, existen covariables que por su naturaleza necesitan ser monitoreadas durante el periodo de seguimiento. Un estudio bastante analizado en la literatura es la del programa de transplantes de corazón en el hospital de Stanford (E. Colosimo, 2006). En este estudio, los pacientes eran aceptados al programa de transplante de corazón previo

análisis de chequeos respectivos de parte de los médicos comprometidos en esta tarea, los pacientes esperaban a algún donador del órgano vital según el orden del diagnóstico por el especialista. El uso de la covariable de que, si recibió corazón o no el paciente tiene dos valores: cero para aquellos pacientes que están esperando el transplante y uno para los que ya recibieron corazón nuevo, por lo que, obviamente ésta covariable es dependiente del tiempo, ya que, el evento de interés es la muerte o fallecimiento del paciente que ha sido seleccionado para esperar el respectivo transplante. Según el caso, habrá paciente que morirán sin recibir el transplante, o sea, recibirán el transplante del corazón aquellos pacientes que sobrevivirán mayor tiempo posible.

Como hemos visto, covariables que alteran sus valores a lo largo del periodo de seguimiento son conocidas como covariables dependientes del tiempo. Tales covariables, cuando se presentan en un estudio, pueden ser considerados en el modelo de Cox generalizado o extendido como sigue:

$$\lambda(t/X_j)=\lambda_0(t)\exp\{X'(t)\beta\} \qquad (4.1)$$

Definida de esta forma, el modelo (4.1) no cumple con la hipótesis de riesgos proporcionales, porque la razón de las funciones de riesgo para el tiempo t y para dos individuos i y j es:

$$\frac{\lambda_i(t)}{\lambda_j(t)} = \exp\{X'_i(t)\beta - X'_j(t)\beta\}$$
$$= \exp\{X'_i(t) - X'_j(t)\}\beta \qquad (4.2)$$

Lo cual se observa que es dependiente del tiempo. En la interpretación de los coeficientes β del modelo, cada componente β_k, para k=1,2,…, p se interpreta como el logaritmo de la razón de riesgo cuyo valor de la k-ésima covariable en el tiempo t difiere de una unidad, cuando los valores de las otras covariables son fijos o constantes en ese instante de tiempo.

El ajuste del modelo de Cox (4.1) es obtenido extendiendo el logaritmo de la función de la verosimilitud parcial dada en (3.9). Esto es:

$$U(\beta)=\sum_{i=1}^{n}\delta_i\left[X_i(t_i)-\frac{\sum_{j\in R(t_i)}X_j(t_i)\exp\left\{X_j'(t_i)\beta\right\}}{\sum_{j\in R(t_i)}\exp\left\{X_j'(t_i)\right\}}\right]=0$$

Donde se debe considerar que las covariables son dependientes del tiempo. En ese aspecto, debemos construir intervalos de confianza y testar hipótesis sobre los coeficientes estimados del modelo.

Ejemplo 10: Supongamos que tenemos un conjunto de datos que presentan eventos de interés, registrado en un periodo de tiempo, o sea, se toma en cuenta tanto el tiempo de inicio y final del estudio, como también se registran los eventos o estados de los casos. Lo cual en lenguaje R se procede así:

> require(survival)

ejemplo1<list(tinicial=c(1,2,5,2,1,7,3,4,8,8),tfinal=c(2,3,6,7,8,9,9,9,14,17), evento=c(1,1,1,1,1,1,1,0,0,0), x=c(1,0,0,1,0,1,1,1,0,0))

> ejemplo1

$tinicial
[1] 1 2 5 2 1 7 3 4 8 8

$tfinal
[1] 2 3 6 7 8 9 9 9 14 17

$evento
[1] 1 1 1 1 1 1 1 0 0 0

$x
[1] 1 0 0 1 0 1 1 1 0 0

> summary(coxph(Surv(tinicial,tfinal,evento)~x,ejemplo1))

Call:
coxph(formula = Surv(tinicial, tfinal, evento) ~ x, data = ejemplo1)

n= 10

Tabla 15: Obtención del coeficiente de la covariable X dependiente del tiempo t.

	coef	exp(coef)	se(coef)	z	p
X	-0.0211	0.98	0.795	-0.0265	0.98

	exp(coef)	exp(-coef)	lower .95	upper .95
	0.98	1.02	0.206	4.65

La salida anterior nos permite verificar la significación del coeficiente de la covariable X para interpretar los resultados obtenidos . El coef=- 0.0211 representa el valor del coeficiente de la covariable X que corresponde a un riesgo relativo=0.98 lo que significa que una X con una característica dada el riesgo es 0.98 veces mayor en comparación con aquello que no tiene tal característica, esto es equivalente a 1/1.02, mientras que 0.795 representa el error estándar que se puede cometer al estimar el valor del coeficiente de la covariable X. La columna z=-0.0265 representa el valor crítico del estadístico de la distribución asintóticamente normal para testar H_A: $\lambda_1(t)/\lambda_0(t)$=Exp(coef) (Zhang D. 2005). El p-valor=0.98 no es significativo respecto al valor estimado del coeficiente de X. Finalmente tenemos el intervalo de confianza [0.206, 4.65] al 95% del riesgo relativo Exp(coef).

Rsquare = 0 (max possible= 0.84).

Likelihood ratio test	= 0 on 1 df, p=0.979
Wald test	= 0 on 1 df, p=0.979
Score (logrank) test	= 0 on 1 df, p=0.979

Tabla de los tres criterios

Si hubiésemos tratado en forma global, o sea, considerando todas las variables juntas y sin tener en cuenta el tiempo de variación adecuada de los eventos, según los resultados de esta tabla podemos decir, que: el sistema de ecuaciones según el método de máximo verosimilitud no es soluble, porque el p-valor= 0.979>0.05. La única solución para el sistema seria 0 según el Test de Wald y, las curvas de supervivencia para dos grupos diferentes de individuos serian iguales, porque, el p-valor del test del Score(lograk) es igual a 0.797, también, los valores de los tres criterios como: Likelihood ratio test, Wald test y Score (logrank) test deben ser lo más altos posibles y no cero o cercanos a ceros. Y finalmente diremos que el problema no se puede modelar mediante el modelo de regresión múltiple, ya que, Rsquare=0.

Como quiera que tenemos el coeficiente estimado, entonces el modelo de Cox estaría dada por:

$$\lambda(t/x)=\lambda_0(t)\exp\{-0.0211x\}$$

para cada valor de x.

4.3.2.- MODELO DE COX ESTRATIFICADO.

El modelo de cox estratificado admite que la forma de la función de riesgo varíe según los estratos o niveles de las covariables. Lo cual significa que: Supongamos que tenemos un predictor, X, por su naturaleza propia, éste X puede ser categorizado en varios niveles, o sea, en subpredictores secundarios, y esto nos

conlleva a que el modelo sea ajustado para cada subpredictor Z. De modo que el modelo de Cox para cada estrato será definido de la manera siguiente:

$$\lambda_{ij}(t) = \lambda(t/X_j , Z=j)=\lambda_{0j}(t)\exp\{X'_{ij}(t)\beta\}, j=1,2,\ldots,m \qquad (4.3)$$

donde m representa el número de estratos y i=1,2,…, n_j , ya que, cada n_j es el número de observaciones en cada estrato. Las funciones de riesgo base $\lambda_{o1,}$ $\lambda_{02,}$…, λ_{0m} son arbitrarios y completamente no correlacionados (Enrico A. Colosimo -2006).
La estratificación no crea ninguna complicación en la estimación del vector de parámetros β. Una función de verosimilitud parcial análoga a (3.7) es construido para cada estrato de modo que se estiman cada unos de los β´s, para luego tener estimado el β general mediante la suma de los logaritmos de las funciones de verosimilitudes parciales, esto es:

$$\ell(\beta)=[\ell_1(\beta)+\ldots+\ell_m(\beta)] \qquad (4.4)$$

Con $\ell_j(\beta)=\ln(L_j(\beta))$ obtenido solamente con los datos de los individuos en el j-ésimo estrato (Enrico A. Colosito-2006). Las derivadas para (4.1) son encontradas por medio de la suma de las derivadas obtenidas para cada estrato, luego $\ell(\beta)$ es maximizada con respecto a β, tal como se hizo en el capítulo anterior.

El modelo de Cox estratificado (4.3) asume que las covariables actúan de modo similar en la función de riesgo de base de cada estrato, o sea, se asume que β es común para todos los estratos (Enrico A. Colosimo-2006) lo cual, ésta suposición debe ser probada. Tal como veremos en la sección 4.4.2.

Ejemplo 11: Supongamos que tenemos un conjunto de datos que presentan eventos de interés, tiempo de ocurrencia del evento, estado o censura. Como covariables se tiene x (conjunto de valores) y sexo (femenino y masculino), la segunda covariable nos perrnite estratificar en dos categorías la conjunto de datos. En lenguaje R tendremos como sigue:

> require(survival)

>ejemplo2<list(tiempo=c(4,3,1,1,2,2,3),estado=c(1,NA,1,0,1,1,0),x=c(0,2,1,1,1,0,0),sexo=c(0,0,0,0,1,1,1))

> ejemplo2

$tiempo
[1] 4 3 1 1 2 2 3

$estado
[1] 1 NA 1 0 1 1 0

$x
[1] 0 2 1 1 1 0 0

$sexo
[1] 0 0 0 0 1 1 1

> summary(coxph(Surv(tiempo,estado)~x+strata(sexo),ejemplo2))

Call:
coxph(formula = Surv(tiempo, estado) ~ x + strata(sexo), data = ejemplo2)

n=6 (1 observations deleted due to missing)

Tabla 16: El Coeficientes de la covariable X de Cox estratificada.

	coef	exp(coef)	se(coef)	z	p
X	1.17	3.22	1.29	0.907	0.36

	exp(coef)	exp(-coef)	lower .95	upper .95
	3.22	0.311	0.258	40.2

La salida anterior nos permite verificar la significación del coeficiente de la covariable X para interpretar los resultados obtenidos. El coef=1.17 representa el valor del coeficiente de la covariable X que corresponde a un riesgo relativo=3.22 lo que significa que una X con una característica dada el riesgo es 3.22 veces mayor en comparación con aquello que no tiene tal característica, esto es equivalente a 1/0.311, mientras que 1.29 representa el error estándar que se puede cometer al estimar el valor del coeficiente de la covariable X. La columna z=0.907 representa el valor crítico del estadístico de la distribución asintóticamente normal para testar H_A: $\lambda_1(t)/\lambda_0(t)=Exp(coef)$ (Zhang D. 2005). El p-valor=0.36 no es significativo respecto al valor estimado del coeficiente de X ya que es mayor que 0.05. Finalmente tenemos el intervalo de confianza [0.258, 40.2] al 95% del riesgo relativo Exp(coef).

Rsquare = 0.135 (max possible= 0.618)

Likelihood ratio test	= 0.87	on 1 df,	p=0.351
Wald test	= 0.82	on 1 df,	p=0.364
Score (logrank) test	= 0.89	on 1 df,	p=0.345

Tabla de los tres criterios

La interpretación de los resultados es análoga a lo que se dio en el ejemplo 10. Por lo tanto, el modelo de Cox estaría dada por:

$$\lambda(t/x)=\lambda_0(t)\exp\{1.17x\}$$

para cada valor de x.

4.4 APLICACIONES.

En el presente trabajo de tesis presentaremos dos aplicaciones que ilustren la forma funcional de los covariables del modelo de Cox. Tal como veremos en las páginas siguientes.

4.4.1 ANALISIS DE LOS DATOS DE PACIENTES HIV.

El estudio de la epidemia de AIDS (SIDA) es un área de intensa investigación, por lo que en este trabajo de tesis presento de manera detallada los factores influyentes que una persona es diagnosticada por infección al *sinusites,* utilizando el Modelo de Cox con covariables dependientes del tiempo, para lo cual, se considera a 112 pacientes, de los cuales, 91 son pacientes HIV positivos y 21 HIV negativos. La clasificación de los pacientes de acuerdo a la infección por HIV es como sigue: HIV seronegativo, HIV seropositivo asintomático, con ARC (AIDS con cierto grado de complejidad) y con AIDS (ver E. Colosimo – 2006). A esta clasificación la denominaremos covariable “grupo de riesgo”. Ahora veamos conceptos respectivos de cada terminología.

En el covariable grupo de riesgo, pacientes HIV seronegativo son aquellos que no poseen el HIV. Pacientes HIV seropositivo asintomático son aquellos que poseen el virus mas no desarrollan un cuadro clínico de AIDS (SIDA) y que presentan un perfil inmunológico estable. Pacientes con ARC son aquellos que presentan baja inmunidad y otros indicadores clínicos que anteceden el cuadro clínico de AIDS y. Pacientes con AIDS son aquellos individuos que ya desarrollaron infecciones oportunistas que definen esta dolencia. Estas definiciones de las terminologías de categorización están basadas según los criterios de CDC de 1987 (Center of Disease Control, 1987-USA). A esta covariable denominaremos “grupo de riesgo”, que es posible que depende del tiempo; porque los pacientes según sus cuadro clínicos pueden mudar de una a otra categoría durante el periodo de estudio u observación. Lógicamente también se consideran otras covariables como el tiempo de inicio y final del estudio, edad, sexo, conteo de células CD4, CD8, etc, etc. Todo ello veremos en la tabla siguiente que demuestra covariables medibles en el estudio de la ocurrencia de sinusites.

Tabla 17: Ocurrencia de sinusites.

Edad del paciente	Medido en años
Sexo del paciente	0- Masculino 1- Femenino
Grupo de Riesgo	1- Paciente HIV seronegativo 2- Paciente HIV seropositivo asintomático 3- Paciente con ARC 4- Paciente con AIDS
Actividad Sexual	1- Homosexual 2- Bisexual 3- Heterosexual
Uso de Droga por Inyección	1- Si 2- No
Uso de Cocaína Por Aspersión	1- Si 2- No

Finalmente tendremos el modelamiento estadístico respectivo, utilizando el lenguaje R. Los datos correspondientes se hallan en la Tabla 1 del anexo.

> require(survival)

> aids<-read.table('c:\\aplicacioncox\\aids.txt',h=T)

> attach(aids)

>ajuste1< coxph (Surv(ti[ti<tf], tf[ti<tf], cens[ti<tf]) ~ ed[ti<tf]+

sex[ti<tf] + factor(grp) [ti<tf] + ats[ti<tf] + ud[ti<tf] + ac[ti<tf])

> summary(ajuste1)

Call:
coxph(formula = Surv(ti[ti < tf], tf[ti < tf], cens[ti < tf]) ~ ed[ti < tf] + sex[ti < tf] +

factor(grp)[ti < tf] + ats[ti < tf] + ud[ti < tf] + ac[ti < tf])

n=96 (27 observations deleted due to missingness)

Tabla 18: Coeficientes de Cox.

	coef	exp(coef)	se(coef)	z	p
ed[ti < tf]	-0.0916	0.912	0.0406	-2.259	0.0240
sex[ti < tf]	0.8380	2.312	0.7019	1.194	0.2300
factor(grp)[ti < tf]2	-0.3335	0.716	1.4220	-0.235	0.8100
factor(grp)[ti < tf]3	2.9485	19.078	1.1702	2.520	0.0120
factor(grp)[ti < tf]4	3.6867	39.914	1.1780	3.130	0.0017
ats[ti < tf]	-0.4339	0.648	0.3770	-1.151	0.2500
ud[ti < tf]	0.3098	1.363	1.0964	0.283	0.7800
ac[ti < tf]	0.3994	1.491	1.4754	0.271	0.7900

	exp(coef)	exp(-coef)	lower .95	upper .95
ed[ti < tf]	0.912	1.0959	0.8427	0.988
sex[ti < tf]	2.312	0.4326	0.5841	9.150
factor(grp)[ti < tf]2	0.716	1.3959	0.0441	11.630
factor(grp)[ti < tf]3	19.078	0.0524	1.9249	189.077
factor(grp)[ti < tf]4	39.914	0.0251	3.9666	401.629
ats[ti < tf]	0.648	1.5433	0.3095	1.357
ud[ti < tf]	1.363	0.7336	0.1590	11.690
ac[ti < tf]	1.491	0.6707	0.0827	26.875

Las interpretaciones de los resultados obtenidos es similar a las interpretaciones dadas en anteriores resultados, lo más importante que podemos observar en esta tabla son los coeficientes del ajuste del modelo de Cox, la covariable grupos de riesgos (factor) que depende del tiempo se observa en la tabla 4.4 en tres oportunidades.

Rsquare= 0.271 (max possible= 0.772)

Likelihood ratio test=	30.4	on 8 df,	p=0.000180
Wald test	= 18.2	on 8 df,	p=0.0196
Score (logrank) test =	29.1	on 8 df,	p=0.000302

Tabla de los tres criterios

Tal como dejimos anteriormente, los resultados obtenidos en la tabla anterior son consistentes, en el sentido de que el sistema correspondiente es soluble, los respectivos coeficientes son diferentes de cero, y las funciones de supervivencia para dos grupos diferentes son proporcionales, finalmente, según el valor de Rsquare podemos decir que el problema no puede ser modelada por una regresión múltiple. Como observamos en la tabla 18, que existen covariables significativos, cuyas estimaciones aparecen en la tabla 19. O sea; con respecto a los valores obtenidos en la tabla 19, podemos decir que los únicos covariables edad y grupos de riesgo son significativos, el resto no son significativos, por lo que, para estas dos covariables significativos obtendremos nuevamente sus estimaciones:

```
> require(survival)
> aids<-read.table('c:\\aplicacioncox\\aids.txt',h=T)
> attach(aids)
```

>ajuste2<coxph(Surv(ti[ti<tf], tf[ti<tf], cens[ti<tf]) ~ ed[ti<tf]+

factor(grp)[ti<tf], method="breslow")

> summary(ajuste2).

Call:
coxph(formula = Surv(ti [ti < tf], tf[ti < tf], cens[ti < tf]) ~ ed[ti < tf] +

factor(grp) [ti < tf], method = "breslow")

n=121 (2 observations deleted due to missingness)

Tabla 19: Coeficientes de Cox para covariables significativos

	coef	exp(coef)	se(coef)	z	p
ed[ti < tf]	-0.0789	0.924	0.0315	-2.505	0.01200
factor(grp)[ti < tf]2	-0.6876	0.503	1.0006	-0.687	0.49000
factor(grp)[ti < tf]3	2.2847	9.822	0.8373	2.729	0.00640
factor(grp)[ti < tf]4	2.6659	14.381	0.7904	3.373	0.00074

	exp(coef)	exp(-coef)	lower .95	upper .95
ed[ti < tf]	0.924	1.0821	0.8688	0.983
factor(grp)[ti < tf]2	0.503	1.9889	0.0707	3.574
factor(grp)[ti < tf]3	9.822	0.1018	1.9034	50.687
factor(grp)[ti < tf]4	14.381	0.0695	3.0548	67.698

En esta tabla tenemos los resultados concernientes a las covariables edad y grupos de riesgo, que son factores influyentes para la ocurrencia de sinusites. Según E. Colosimo que, a cada 10 años en la edad de los pacientes, el riesgo de desarrollar sinusites disminuye en 54% (1-exp(-0.0789*10)=0.54), lo que indica que los pacientes más jóvenes son los más propenso a esta infección. Por otro lado, podemos deducir de la tabla anterior que, los pacientes del grupo HIV seropositivo asintomático no defieren significativamente del riesgo de los pacientes HIV seronegativo. Mientras que, el grupo ARC que desarrolla sinusites es exp(2.2847)=9.822 veces mayor el riesgo del grupo HIV seronegativo. Para el grupo con AIDS, el riesgo de desarrollar sinusites es 14.381 veces mayor el riesgo del grupo HIV seronegativo. Por otro lado, la precisión de las estimativas asociadas a estas dos últimas razones de riesgo es bastante reducida, como puede ser observada por la gran amplitud de sus respectivos intervalos de clase. En la siguiente tabla tendremos la información de que los valores obtenidos en la tabla anterior son consistentes, tal como indican los respectivos p-valores.

Rsquare= 0.271 (max possible= 0.832)

Likelihood ratio test = 38.3 on 4 df, p=9.86e-08

Wald test = 25.8 on 4 df, p=3.54e-05

Score (logrank) test = 38.5 on 4 df, p=8.99e-08

Tabla de los tres criterios

Según ésta tabla podemos decir que los valores obtenidos en la tabla anterior son también consistentes, por lo tanto el modelo existe.

> cox.zph(ajuste2,transform="identity")

	rho	chisq	p

ed[ti < tf]	-0.0752	0.1456	0.703
factor(grp)[ti < tf]2	0.1604	0.6755	0.411
factor(grp)[ti < tf]3	0.0350	0.0322	0.858
factor(grp)[ti < tf]4	0.1107	0.3240	0.569
GLOBAL	NA	1.1665	0.884

Según ésta tabla podemos decir que: la determinación adecuada del covariable grupo como dependiente de tiempo, no se viola el supuesto de riesgos proporcionales al nivel del confianza del 5%, tal como podemos apreciar en los siguientes figuras de residuos de Schoenfeld.

> **par(mfrow=c(1,4))**

> **plot(cox.zph(ajuste2))**

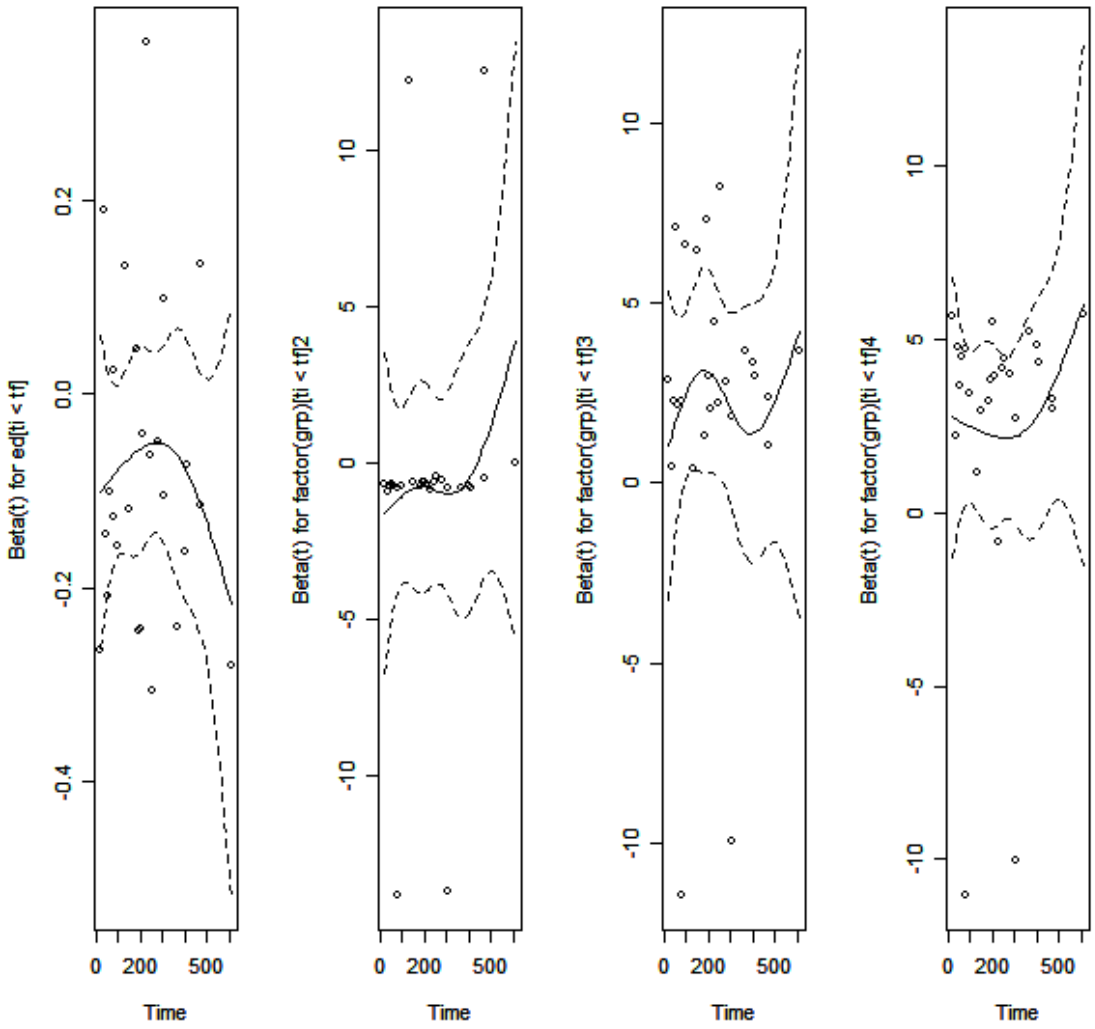

Figura 20: Residuos estandarizados de Schoenfeld versus los tiempos para las covariables considerados en el modelo de Cox dependiente de tiempo.

Observamos en la figura 20 que los residuos de Schoenfeld varían de manera aleatoria en las cercanías de cero y no presentan algún patrón o seguimiento extraño. Por lo tanto, se garantiza el cumplimiento de la hipótesis de riesgo proporcional para los dos covariables significativos dependientes del tiempo.

4.4.2 MODELO DE COX ESTRATIFICADO CON DATOS DE LEUCEMIA.

El modelo de Cox es estratificado en que los individuos sean separados en dos diferentes estratos, de acuerdo con las categorías de la covariable LEUINIC (verTabla 2 del Anexo), y el modelo correspondientes es:

$$\lambda_{ij}(t) = \lambda(t/X_j, Z=j)=\lambda_{0j}(t)\exp\{X'_{ij}(t)\beta\}, j=1,2; i=1,2,\ldots,n_j; \qquad (4.5)$$

donde n_j representa el número de individuos en cada estrato; para tal efecto, definimos los estratos correspondientes de la forma siguientes; LEUINIC≤75000 mm^3 e aquellos con LEUINIC >75000 mm^3, para ajustar el modelo (4.5), asumimos que el vector β es común en los dos estratos, ésta conjetura será testada en la tabla 21, para lo cual, la Tabla 2 del anexo será transformada según los comandos siguientes del lenguaje R, cuyo resultado de la transformación se observa en la Tabla 3 del anexo.

```
> require(survival)
> leuc<-read.table('c:\\aplicacioncox\\leucemia.txt',h=T)
> attach(leuc)
> edadc<-ifelse(edad>96,1,0)
> leuinicc<-ifelse(leuinic>75,1,0)
> zpesoc<-ifelse(zpeso>-2,1,0)
> zestc<-ifelse(zest>-2,1,0)
> pasc<-ifelse(pas>0.05,1,0)
> vacc<-ifelse(vac>15,1,0)
> riesgoc<-ifelse(riesgo>1.7,1,0)
> r6c<-r6
> leucc<-as.data.frame(cbind(leuinicc,tiempo,estado,edadc,zpesoc,zestc,
   pasc,vacc,riesgoc,r6c))
```

> leucc

> detach(leuc)

> attach(leucc)

>ajuste<coxph(Surv(tiempo,estado)~edadc+zpesoc+pasc+vacc+strata(leuinic c),data=leucc, x=T,method="breslow")

> summary(ajuste)

Call:
coxph(formula = Surv(tiempo, estado) ~ edadc + zpesoc + pasc + vacc +
strata(leuinicc),
data = leucc, method = "breslow", x = T)

n= 103

Tabla 20: Probabilidades significativas de las covariables. (EN LA VERSION PDF, PASAR A LA SIGUIENTE PAGINA)

	coef	exp(coef)	se(coef)	z	p
edadc	0.737	2.090	0.394	1.87	6.1e-02
zpesoc	-2.024	0.132	0.497	-4.08	4.6e-05
pasc	-0.421	0.656	0.396	-1.06	2.9e-01
vacc	1.120	3.064	0.414	2.71	6.8e-03

	exp(coef)	exp(-coef)	lower .95	upper .95
edadc	2.090	0.478	0.9666	4.520
zpesoc	0.132	7.572	0.0499	0.350

pasc	0.656	1.524	0.3023	1.425
vacc	3.064	0.326	1.3617	6.893

Rsquare = 0.182 (max possible= 0.921)

Likelihood ratio test	= 20.7 on 4 df, p=0.000365
Wald test	= 23.4 on 4 df, p=0.000107
Score (logrank) test	= 29.4 on 4 df, p=6.47e-06

La tabla de los tres criterios

En esta tabla observamos los valores estimados del vector β. Según el lenguaje R tienen probabilidades significativas las covariables EDAD, ZPESO, PAS y VAC. Por lo que, el modelo correspondiente es:

$$\lambda_{ij}(t)=\lambda(t/X_j, Z=j)=\lambda_{0j}(t)\exp\{0.737edad_j-2.024zpeso_j-0.421pas_j+1.120vac_j\} \quad (4.6)$$

para j=1,2; y éstas son las covariables más influyentes en el problema de leucemia. Ahora, probaremos que el vector β es común en los dos estratos, por medio del siguiente comando en lenguaje R.

> cox.zph(ajuste,transform="identity")

Tabla 21: Tabla de rho, Chi-cuadrado y p-valor

	rho	chisq	p
edadc	-0.1089	0.52787	0.468
zpesoc	0.0923	0.30499	0.581

pasc	0.0601	0.13830	0.710
vacc	-0.0139	0.00853	0.926
GLOBAL	NA	0.83075	0.934

En esta tabla observamos no hay evidencia de que los β's sean distintos entre los estratos, ya que, los correspondientes probabilidades no son significativos, o sea, eso implica que se acepta la supuesta hipótesis de que los β's no sean diferentes en cada estrato.

> par(mfrow=c(1,4))

> plot(cox.zph(ajuste))

El resultado es el siguiente grupo de gráficos, conocidos como gráficos de los residuos estandarizados de Schoenfeld.

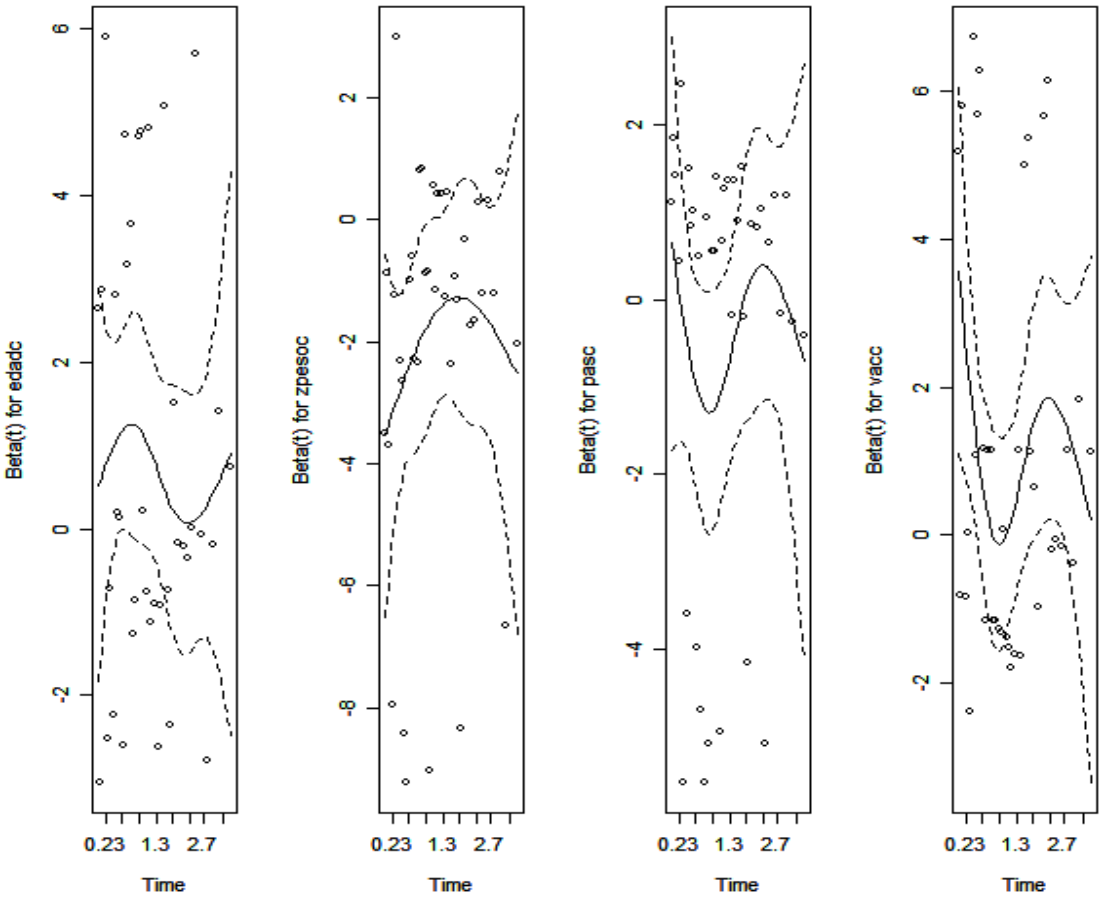

Figura 21: Residuos estandarizados de Schoenfeld versus los tiempos para las covariables considerados en el modelo de Cox estratificado. (JUNTAR CON LA FIGURA CORRESPONDIENTE)

En la figura 21 observamos que los residuos de Schoenfeld varían aleatoriamente alrededor de 0, en consecuencia, podemos garantizar el cumplimiento de la hipótesis de riesgo proporcionales para los cuatro variables que aparecen en el modelo (4.6), o sea, estos residuos de la variables preponderantes no presenta tendencia alguna o algún patrón extraño.

CONCLUSIONES

1. El modelo de regresión de Cox es el modelo que relaciona la función de riesgo para cada individuo como producto de dos funciones; una que representa el efecto de las covariables y otro el de tiempo. Esto implica que la estructura de este modelo impone proporcionalidad entre funciones de riesgo para diferentes niveles de covariables; o sea, para dos individuos diferentes, la razón de dos funciones de riesgo es constante respecto al tiempo y. Los coeficientes correspondientes son estimadas a partir de las observaciones muestrales los cuales nos permiten medir los efectos de las covariables en la función del riesgo.

2. El modelo de Cox es adecuado para datos en que la suposición de riesgos proporcionales es válida; por lo tanto, ésta suposición debe verificarse si verdaderamente el modelo construido se ajusta a los datos obtenidos, con esa finalidad se utiliza el análisis de residuos de Schoenfeld.

3. Ocurre casos en que la suposición de riesgos proporcionales del modelo de Cox no se cumple, lo cual significa que el modelo obtenido no se adecua a los datos del análisis de supervivencia, porque las covariables son dependientes del tiempo o que el problema necesita estudiar por estratos de modo que, en cada estrato cumpla la hipótesis fundamental de riesgos proporcionales.

4. Cuando las covariables son dependientes de tiempo, la suposición fundamental de riesgos proporcionales del modelo de Cox no se cumple; porque los datos son observaciones longitudinales tiempo-dependientes, en consecuencia, la razón de riesgos proporcionales de dos niveles resultan ser funcionales del tiempo. En tal caso, los datos obtenidos según su propia naturaleza serán estratificadas adecuadamente para garantizar el cumplimiento de la suposición de riesgos proporcionales del modelo de Cox y, la buena adecuación del modelo serán probadas por medio de los criterios de Likelihood ratio test, Wald test y Score Test, corroborada por las gráficas de residuos de Schoenfeld mediante la no presencia de un patrón extraño.

BIBLIOGRAFIA

1. ABAURREA J., CEBRIAN A. ; Fiabilidad y Análisis de Supervivencia. Departamento de Métodos Estadísticos - Universidad de Zaragoza, España 2004.

2. ALLISON PAUL D. Survival Analysis Using SAS System: am practical guide cary, SAS Institute Inc. USA 1995.

3. BERRENDERO JOSE R.; Descripción de los Datos con el Entorno R. Dpto. de Matemáticas de la Universidad Autónoma de Madrid- 2005.

4. BRAVO ANTONIO; Análisis de Datos de Sobrevivencia: Análisis Paramétrico, el Estimador de Kaplan-Meir. Tesis de Grado, UNMSM 1990.

5. COLLET DAVID; Modelling Survival Data in Medical Research. London, Champamn & Hall, 1994.

6. COX D. R. ; Regression Models and Life Tables. JRSS Series B, Vol 34 1972.

7. ENRICO ANTONIO COLOSIMO, SUELE RUIZ GIOLO; Análise de Sobrevivência Aplicada, ABE-Projeto Fisher-Brasil 2006.

8. COX D. R. AND OAKES D. , Analysis of Survival Data, Edit. Chapman and Hall London 1984.

9. LAWLESS, J. F.; Statistical Models and Methods for Lifetime Data, Edit. Jhon Wiley & Sons, USA 1982.

10. LEE ELISA T., Statistical Methods for survival Data Análisis, Lifetime Learning

Publications Belmont, California 1980.

11. LOUZADA NETO FRANCISCO, MAZUCHELI JOSMAR, ACHCAR JORGE ALBERTO, Análise de Sobrevivënci e Confiabidade, IMCA 2002.

12. MEI-CHENG WANG, Summary Notes for Survival Analysis. Departament of Biostatistics Johns Hopkins University 2005.

13. MORENO VICTOR, Análisis de la Supervivencia, 2004.

14. MONTGOMERY-PECK-VINING, Introducción al Análisis de Regresión Lineal, CECSA 3ª Edición 2004 Mexico.

15. MILLER RUPERT G. Survival Analysis, John Wiley and Sons, Inc. USA 1976.

16. RIVAS LÓPEZ, MARIA JESÚS y LÓPEZ FIDALGO, JESÚS; Análisis de Supervivencia. Edit. La Muralla, Madrid , España 2000.

17. STEFANESCU CATALINA, AND MEHROTRA DEVAN. Cox Model Versus Generalized Logrank Test for time-to-event Data with Ties. Journal of the Royal Statistical Society, 2003.

18. THERNEAU TERRY M. AND GRAMBSCH PATRICIA M., Modeling Survival Data: Extending Cox Model, Springer 2000.

19. YUDO WADA CICILIA, Análisis de Supervivencia, Universidad Nacional Mayor de San Marcos- Facultad de Ciencias Matemáticas, Lima Perú 1999.

20. ZAMORA MUÑOZ SALVADOR, SALAZAR MARTINEZ EDUARDO Y LAZCANO PONCE EDUARDO. Análisis de Supervivencia. Aplicación en una Muestra de Mujeres con Cáncer Cervical en México. Centro de Investigación en

Salud Poblacional, Instituto Nacional de Salud Pública (INSP). Cuernavaca, Morelos, México 2000 . Vol 42, número 3.

21. ZHANG DAOWEN, Análisis of Survival Data, Departament of Statistic North Carolina State University, USA 2005.

ANEXO

Tabla 1: Datos utilizados en el estudio sobre paciente con HIV (pág.81)

pac	ed	sex	grp	ti	tf	cens	cd4	cd8	ats	ud	ac
1	31	0	4	0	0	1	NA	NA	3	2	2
2	22	1	2	0	378	0	132	715	3	2	2
3	32	0	4	0	84	1	75	315	3	2	2
4	36	0	2	0	109	0	NA	NA	3	2	2
5	34	0	2	0	134	1	NA	NA	NA	NA	NA
6	29	0	2	0	338	0	NA	NA	1	2	2
7	29	1	3	0	311	0	73	590	3	2	2
8	22	0	4	0	0	0	58	775	2	2	2
9	38	0	4	0	182	1	NA	NA	1	2	2
10	30	1	1	0	77	1	NA	NA	3	2	2
11	30	1	1	0	184	0	NA	NA	3	2	2
12	33	0	2	0	543	0	310	870	1	2	2
13	35	1	1	0	286	0	NA	NA	3	2	2
14	41	0	4	0	470	1	235	746	1	2	2
15	31	0	4	0	407	1	NA	NA	NA	NA	NA
16	48	1	3	0	231	1	NA	NA	3	2	2
17	31	0	2	0	205	0	420	725	NA	NA	NA
18	21	1	1	0	637	0	NA	NA	3	2	2
19	22	1	1	0	345	0	NA	NA	3	2	2
20	32	0	1	0	638	0	NA	NA	1	2	2
21	37	1	1	0	292	0	NA	NA	3	2	2
22	25	0	1	0	294	0	NA	NA	NA	NA	NA
23	NA	0	2	0	471.5	0	NA	NA	1	2	2
23	NA	0	4	471.5	507	0	NA	NA	1	2	2
24	34	1	3	0	141.5	0	200	3	3	2	2
24	34	1	4	141.5	244.5	1	5	200	3	2	2
25	31	0	4	0	49	1	NA	NA	1	1	2
26	27	0	4	0	511	0	NA	NA	3	1	1
27	20	0	2	0	498	0	210	606	3	2	2
27	20	0	4	498	611	1	210	606	3	2	2
28	27	0	2	0	308	0	NA	NA	NA	NA	NA
28	27	0	4	308	371	1	NA	NA	NA	NA	NA
29	31	0	4	0	681	0	30	700	3	2	2
30	48	1	2	0	703	0	610	585	3	2	2
31	41	1	1	0	660	0	417	190	3	2	2
32	23	0	1	0	661	0	527	320	2	2	2
33	22	1	2	0	492	0	NA	NA	3	1	1
34	40	0	3	0	42	0	48	885	2	2	2

34	40	0	4	42	583	0	48	885	2	2	2
35	53	0	2	0	276.5	0	200	475	3	2	2
35	53	0	3	276.5	611	0	200	475	3	2	2
36	44	0	4	0	35	1	5	250	2	2	2
37	25	1	1	0	562	0	NA	NA	3	2	2
38	23	0	2	0	665	0	458	420	NA	NA	NA
39	32	0	4	0	294	0	53	160	2	1	2
40	20	0	2	0	0	1	278	865	1	2	2
41	32	1	2	0	644	0	218	400	3	2	2
42	23	0	2	0	266	0	360	850	1	1	1
43	25	0	2	0	273	0	55	295	2	2	2
44	47	1	4	0	525	0	250	485	3	2	2
45	52	1	2	0	143.5	0	130	840	3	2	2
45	52	1	4	143.5	619	0	130	840	3	2	2
46	32	0	2	0	94.5	0	173	1070	3	2	2
46	32	0	3	94.5	617	0	173	1070	3	2	2
47	26	0	2	0	634	0	NA	NA	NA	NA	NA
48	30	0	3	0	274	0	12.5	235	NA	NA	NA
48	30	0	4	274	315	0	12.5	235	NA	NA	NA
49	37	1	2	0	609	0	NA	NA	3	2	2
50	40	0	2	0	598	0	373	420	3	2	2
51	35	0	2	0	548	0	305	715	1	2	2
52	26	1	2	0	589	0	295	1145	3	2	2
53	26	0	2	527	0	268	405	2	2	2	2
54	28	1	2	0	597	0	187	255	3	2	2
55	24	0	2	0	323	0	NA	NA	NA	NA	NA
56	35	1	4	0	415	1	8	140	3	2	2
57	24	1	2	0	469.5	1	185	670	3	2	2
58	38	0	1	0	330	0	507	550	1	2	2
59	20	0	1	0	499	0	NA	NA	1	2	2
60	27	0	4	0	199.5	1	NA	NA	NA	NA	NA
61	23	0	2	0	425.5	0	213.5	1055	3	2	2
61	23	0	3	425.5	478	0	213	1055	3	2	2
62	28	0	3	0	101.5	1	NA	NA	2	1	1
63	55	1	4	0	0	1	135	595	3	2	2
64	40	0	2	0	42	0	50	480	2	2	2
64	40	0	3	42	140	0	50	480	2	2	2
64	40	0	4	140	310.5	1	50	480	2	2	2
65	42	0	4	0	455	0	17.5	340	3	2	2
66	19	0	2	0	444	0	900	1085	1	2	2
67	34	0	4	0	98	0	5	227	1	2	2
68	29	0	3	0	204	0	NA	NA	1	2	2
68	29	0	4	204	248	0	NA	NA	1	2	2
69	29	0	3	0	147	1	327	1505	NA	NA	NA
70	49	0	1	0	283	0	NA	NA	3	2	2

71	50	0	4	0	0	1	67.5	950	3	2	2
72	37	0	1	0	351	0	NA	NA	1	2	2
73	35	0	2	0	365	0	275	1210	1	2	2
74	27	1	2	0	329	0	427	1315	3	2	2
75	26	0	3	0	52.5	1	72	430	2	2	2
76	33	0	4	0	59.5	1	12.5	85	3	2	2
77	22	0	1	0	367	0	NA	NA	1	2	2
78	37	0	3	0	0	1	85	1215	2	2	2
79	47	0	2	0	371	0	127	790	1	2	2
80	25	0	1	0	306.5	1	NA	NA	NA	NA	NA
81	23	0	2	0	343	0	NA	NA	NA	NA	NA
82	35	1	4	0	278.5	1	NA	NA	3	2	2
83	34	0	4	0	325	0	20	97	NA	NA	NA
84	26	0	2	0	330	0	243	705	NA	NA	NA
85	35	0	1	0	260	0	NA	NA	2	2	2
86	24	0	1	0	304	0	NA	NA	2	1	1
87	31	0	3	0	158.5	0	NA	NA	3	2	2
87	31	0	4	158.5	267	0	NA	NA	3	2	2
88	32	1	2	0	297	0	563	975	3	2	2
89	36	0	2	0	297	0	327	525	NA	NA	NA
90	53	0	3	0	275	0	38	290	1	2	2
91	31	1	4	0	13	0	68	425	3	1	1
92	22	0	2	0	125.5	0	370	905	1	2	2
92	22	0	3	125.5	254.5	1	370	905	1	2	2
93	40	0	3	0	43	0	NA	NA	1	2	2
93	40	0	4	43	259	0	NA	NA	1	2	2
94	37	0	2	0	295	0	290	805	1	2	2
95	45	0	3	0	303	0	35	410	1	2	2
96	22	1	2	0	290	0	368	625	3	2	2
97	38	0	4	0	0	1	18	233	2	2	2
98	27	1	2	0	295	0	NA	NA	3	2	2
99	35	0	4	0	209.5	1	670	900	3	2	2
100	31	0	2	0	139	0	50	560	2	1	2
100	31	0	4	139	283	0	50	560	2	1	2
101	23	0	2	0	242	0	413	810	NA	NA	NA
102	49	0	3	0	125.5	0	42.5	320	2	2	2
102	49	0	4	125.5	295	0	42.5	320	2	2	2
103	27	0	2	0	247	0	437	850	3	2	2
104	38	0	4	0	0	1	20	155	NA	NA	NA
105	25	0	1	0	267	0	290	173	NA	NA	NA
106	40	0	2	0	0	1	270	1920	1	2	2
107	26	0	4	0	19	1	193	770	NA	NA	NA
108	59	0	1	0	269	0	NA	NA	NA	NA	NA
109	30	1	2	0	130.5	0	277	1530	3	2	2
109	30	1	3	130.5	247	0	277	1530	3	2	2

109	30	1	4	247	296	0	277	1530	3	2	2
110	42	0	2	0	247	0	257	510	2	2	2
111	24	0	2	0	86.5	0	57	170	NA	NA	NA
111	24	0	3	86.5	192.5	1	57	170	NA	NA	NA
112	24	0	2	0	226	0	NA	NA	NA	NA	NA

Fuente: Enrique A. Colosimo UFMG Brasil-2006.

pac=paciente, ed=edad (años), sex=sexo (0=masculino, 1=femenino), grp=grupo de riesgo (1=seronegativo, 2=seropositivo asintomático, 3=ARC, 4=aids), ti=tiempo inicial en el grupo, tf=tiempo final en el grupo, cens=censurado (0=censurado, 1=falla), cd4=conteo de CD4, cd8=conteo de CD8, ats=actividad sexual (1=homosexual, 2=bisexual, 3=heterosexual), ud=uso de droga por inyección (1=si, 2=no), ac=aspira cocaína (1=si, 2=no), NA=valor no observado (missing).

Tabla 2: Datos utilizados en el problema de Leucemia (pág. 84)

leuinic	tiempo	cens	edad	zpeso	zest	pas	vac	riesgo	r6
380	1.76	1	60.52	-0.97	-0.48	0.1	5.7	1.58	1
328	0.26	1	68.04	0.36	1.44	0.6	1.5	1.64	0
84.7	0.129	1	156.93	-1.84	-2.17	0.6	20.4	1.26	1
2.9	3.639	1	92.91	-1.06	-0.69	0.7	1.5	0.96	1
400	4.331	0	156.98	-0.84	-0.82	13.7	1	1.32	1
64	4.252	0	69.62	-0.2	-0.19	2.3	2	1.4	1
13.2	0.687	1	79.08	0.02	-2.29	0.3	2	1.52	1
50	0.003	0	112.43	-1.86	-2.42	0	2.7	1.72	1
34.9	2.07	1	47.97	0.15	0.49	4.5	0	1.8	1
68.3	0.709	1	37.91	-0.21	1.27	0	2.1	1.2	1
1	3.466	1	95.21	-2.08	0	43.7	6	0.54	1
24	0.616	1	146.37	-0.49	0.13	0.1	0	1.24	1
140	3.896	0	56.77	-0.07	-1.8	1.2	1.6	1.79	1
5	3.83	0	32.33	-0.32	0.23	0.7	15	0.85	1
49	0.454	1	29.57	0.27	1.11	0	0	2.26	1
68	2.65	1	79.74	-0.66	-1.15	0.1	1	1.3	1
176	3.915	0	160.13	-0.33	-0.98	0.7	4.5	1.06	1
1.6	2.333	1	65.25	2.78	-0.47	0	5	0.6	1
44.6	3.754	0	57.79	0.43	0.19	6.2	13	1.6	1
23.3	1.27	1	69.88	0.57	-0.74	0.7	6.8	1.4	1
6.4	3.704	0	41.59	1.78	1.04	0.2	9.6	0.82	1
15	0.383	0	60.06	-0.51	-0.75	0.5	1.5	1.54	0
96	3.578	0	85.09	-0.74	-1.1	0.3	1.6	1.8	0
4.9	2.902	1	87.06	0.27	0.38	7.8	14	0.72	1
58.2	3.518	0	36.86	-0.17	0.64	0.3	1	1.14	1
6.6	3.485	0	35.94	-0.88	-0.23	0.9	12.8	1.45	1
11.1	2.119	1	86.57	-1.43	-0.33	3.7	24.5	1.16	1
7.5	2.502	1	176.56	-0.84	0.52	0.5	4.3	1.06	0
4.8	3.425	0	70.28	-0.79	-0.36	11.2	1.5	1.3	1
11.7	3.403	0	130.14	0.04	-0.05	0.3	5.3	1.22	1
60	0.715	1	100.34	-0.08	-0.72	0.2	6	1.6	0
3.4	3.198	0	24.41	0.94	2.2	0	5.6	0.9	1
8.7	3.11	0	70.44	-0.31	-1.1	1.2	8.5	0.95	1
2.9	3.209	0	49.45	-0.21	1.6	0.4	12.2	0.58	1
14.8	0.268	0	31.97	0.52	-0.26	0.5	1	1.58	1
168	0.025	1	107.99	0.2	1.38	1.8	16.2	1.36	0
69.8	3.014	0	90.61	-1.91	0.26	0	4.8	1.66	1
123	0.46	1	8.51	-1.44	-0.65	0	15.7	1.6	1
121	2.762	1	38.44	-0.15	0.09	0.4	2.3	1.52	1
86	1.306	1	55.06	0.06	-2.72	0.3	5	1.56	1

3.1	2.053	1	51.52	-3.66	-2.1	0	1.1	0.1	1
74	3.006	0	60.32	3.42	0.63	0	44.1	2	1
13.6	2.861	0	72.48	0.25	-1.09	0	0	1.16	1
1.2	1.227	1	57.86	0.24	1.33	0.3	0.1	0.64	1
58.7	2.264	1	36.96	1.48	0.62	21.9	87.7	1.6	1
62.2	0.841	1	82.89	-2.26	-2.77	0.3	0	1.14	1
4.8	0.917	1	124.81	0.11	0.35	0.3	1	1.08	1
51	2.765	0	61.7	-0.46	0.16	35.8	9	1.28	1
30.1	2.738	0	77.73	-1.1	-0.77	6	7.2	1.18	1
8.7	2.757	0	94.29	-1.43	-0.04	0.1	2.5	1.26	1
3.9	2.639	0	90.22	1.07	2.98	51.7	11.7	0.58	1
2.9	0.736	1	99.48	-1.49	-0.98	1.8	0.7	1.42	1
81	0.63	1	132.4	-1.5	-1.85	0	1	2.1	1
8.1	2.464	0	46.16	-1.44	-0.38	39.5	50.3	0.7	1
5.8	2.428	0	42.12	1.04	0.45	4.5	0.2	1.25	1
4.9	1.443	1	105.53	0.14	-0.19	1	55.7	2.1	1
340.8	0.654	1	132.47	-0.56	-0.67	0	3.6	1.72	1
23	2.355	0	153.13	1.59	0.64	0.9	1.8	1.38	1
27.2	2.278	0	84.34	-0.51	0.23	0	1.5	1.3	1
40.8	0.843	1	16.56	-1.63	-0.34	0	0.1	1.74	1
22.5	2.344	0	48.07	0.01	-0.41	1	2.1	1.68	1
13.2	2.171	0	54.93	0.07	-1.14	29.9	0	1.5	1
9.7	2.133	0	18.96	0.87	0.68	27.4	11.9	1.65	1
32.6	2.22	0	29.21	-0.66	0.05	3.3	15.9	1.2	1
8.1	1.322	1	39.69	0.12	0.63	57.4	5.6	1.26	1
113	0.594	1	60.85	0.43	0.71	0	2	2.28	1
19.4	1.96	0	94.46	1.26	-0.95	0	5	1.78	1
4.2	1.927	0	43.93	-0.56	-0.09	4.7	0	0.95	1
10.8	1.832	0	21.98	-0.7	-0.22	49.4	11.5	1	1
69.3	1.941	0	133.13	-0.71	-1.01	0	12	1.8	1
120	0.099	1	90.25	-0.73	-1.43	0.3	0	1.21	1
5.3	1.714	0	33.25	0.53	1.4	0	0	0.92	1
80.5	0.151	0	137.46	-1.21	-0.01	1.8	1.6	1.3	1
4.5	1.697	0	79.67	0.28	0.1	22	0.1	0.89	1
4	1.692	0	115.25	-0.48	0.45	45.7	39.5	0.62	1
1.2	0.214	1	169.07	-2.32	-1.95	3.3	2	0.88	1
69.4	1.624	0	52.96	-0.93	-1.08	37.1	17.9	1.52	1
4.1	1.566	1	75.17	-1.02	0.08	8.4	19.7	1.5	1
4.2	1.528	0	48.99	-1.56	-0.36	0.3	1	0.76	1
61	1.52	0	62	0.71	-0.99	0.4	0	1.06	1
620	0.487	1	115.22	2.06	1.51	0.6	1	2.7	0
2.1	1.481	0	81.64	0.04	0.48	83.1	64.4	0.78	1
107.5	1.41	0	105	-0.38	-0.15	40.5	5	1.4	1
11.4	0.003	0	63.08	-1.65	-0.34	0.5	1.5	1.28	1
1.3	1.259	0	98.3	-1.03	-0.55	21.3	68.7	1.1	1

1.4	1.205	0	49.68	-1.23	-2.55	0.7	0.3	0.78	1
65.4	1.18	0	79.11	0.31	1.01	0.4	10.1	1.4	1
9.7	0.572	1	66.76	-2.46	-3.05	71.4	19.7	1.12	1
3.8	1.12	0	97.18	-0.33	-0.16	5.7	4	1.7	1
3.6	1.103	0	20.47	-0.93	-0.42	52.3	8.2	1.42	1
31.7	1.065	0	141.54	-1.55	-0.59	4.2	6.5	1.12	1
6	0.498	1	23.69	-2.72	-2.21	1.5	40.5	0.92	1
9	0.991	0	52.27	-0.91	-0.35	1.5	4.9	1.2	1
17.1	0.991	0	74.55	-1.86	-1.18	7.9	3.1	1	1
26.1	0.994	0	86.7	-0.16	-0.34	6.6	5.7	0.88	1
112	0.898	0	57.43	-0.12	-0.99	3	1.7	1.7	1
7	0.969	0	37.91	-1.79	-1.61	0.9	1.1	1.6	1
5.9	0.895	0	90.09	-1.06	-0.96	0.2	2	0.85	1
102	0.893	0	56.54	0.35	-0.35	53	14.2	1.24	1
24.4	0.701	0	72.18	-2.68	-3.7	2.9	3.2	1.46	1
14.1	0.81	0	21.59	-0.82	-0.19	13.3	12.7	1.2	1
5.6	0.742	0	122.58	0	0.34	0.7	2.5	0.72	1
6.5	0.758	0	88.25	-0.97	-0.11	6.3	1.7	0.75	1

Fuente: Enrico A. Colosimo, UFMG, Brasil-2006.

leuini=en 1000 leucocitos/mm^3(conteo de leucocitos iniciales en la sangre periférico); tiempo=respuesta en años; cens= 1 si falla y 0 si censura; edad=en meses; zpeso=peso estandarizado por la edad y sexo; zest=estatura estandarizada por la edad y sexo; pas en %, vac en %, riesgo= factor riesgo en % y r6=1 si ocurre.

Tabla 3.- Valores obtenidos según los comandos correspondientes del lenguaje R para utilizar en la aplicación del modelo de Cox estratificado con datos de leucemia.

Nº	leuinicc	tiempo	estado	edadc	zpesoc	zestc	pasc	vacc	riesgoc	r6c
1	1	1.760	1	0	1	1	1	0	0	1
2	1	0.260	1	0	1	1	1	0	0	0
3	1	0.129	1	1	1	0	1	1	0	1
4	0	3.639	1	0	1	1	1	0	0	1
5	1	4.331	0	1	1	1	1	0	0	1
6	0	4.252	0	0	1	1	1	0	0	1
7	0	0.687	1	0	1	0	1	0	0	1
8	0	0.003	0	1	1	0	0	0	1	1
9	0	2.070	1	0	1	1	1	0	1	1
10	0	0.709	1	0	1	1	0	0	0	1
11	0	3.466	1	0	0	1	1	0	0	1
12	0	0.616	1	1	1	1	1	0	0	1
13	1	3.896	0	0	1	1	1	0	1	1
14	0	3.830	0	0	1	1	1	0	0	1
15	0	0.454	1	0	1	1	0	0	1	1
16	0	2.650	1	0	1	1	1	0	0	1
17	1	3.915	0	1	1	1	1	0	0	1
18	0	2.333	1	0	1	1	0	0	0	1
19	0	3.754	0	0	1	1	1	0	0	1
20	0	1.270	1	0	1	1	1	0	0	1
21	0	3.704	0	0	1	1	1	0	0	1
22	0	0.383	0	0	1	1	1	0	0	0
23	1	3.578	0	0	1	1	1	0	1	0
24	0	2.902	1	0	1	1	1	0	0	1
25	0	3.518	0	0	1	1	1	0	0	1
26	0	3.485	0	0	1	1	1	0	0	1
27	0	2.119	1	0	1	1	1	1	0	1
28	0	2.502	1	1	1	1	1	0	0	0
29	0	3.425	0	0	1	1	1	0	0	1
30	0	3.403	0	1	1	1	1	0	0	1
31	0	0.715	1	1	1	1	1	0	0	0
32	0	3.198	0	0	1	1	0	0	0	1
33	0	3.110	0	0	1	1	1	0	0	1
34	0	3.209	0	0	1	1	1	0	0	1
35	0	0.268	0	0	1	1	1	0	0	1
36	1	0.025	1	1	1	1	1	1	0	0
37	0	3.014	0	0	1	1	0	0	0	1
38	1	0.460	1	0	1	1	0	1	0	1

39	1	2.762	1	0	1	1	1	0	0	1
40	1	1.306	1	0	1	0	1	0	0	1
41	0	2.053	1	0	0	0	0	0	0	1
42	0	3.006	0	0	1	1	0	1	1	1
43	0	2.861	0	0	1	1	0	0	0	1
44	0	1.227	1	0	1	1	1	0	0	1
45	0	2.264	1	0	1	1	1	1	0	1
46	0	0.841	1	0	0	0	1	0	0	1
47	0	0.917	1	1	1	1	1	0	0	1
48	0	2.765	0	0	1	1	1	0	0	1
49	0	2.738	0	0	1	1	1	0	0	1
50	0	2.757	0	0	1	1	1	0	0	1
51	0	2.639	0	0	1	1	1	0	0	1
52	0	0.736	1	1	1	1	1	0	0	1
53	1	0.630	1	1	1	1	0	0	1	1
54	0	2.464	0	0	1	1	1	1	0	1
55	0	2.428	0	0	1	1	1	0	0	1
56	0	1.443	1	1	1	1	1	1	1	1
57	1	0.654	1	1	1	1	0	0	1	1
58	0	2.355	0	1	1	1	1	0	0	1
59	0	2.278	0	0	1	1	0	0	0	1
60	0	0.843	1	0	1	1	0	0	1	1
61	0	2.344	0	0	1	1	1	0	0	1
62	0	2.171	0	0	1	1	1	0	0	1
63	0	2.133	0	0	1	1	1	0	0	1
64	0	2.220	0	0	1	1	1	1	0	1
65	0	1.322	1	0	1	1	1	0	0	1
66	1	0.594	1	0	1	1	0	0	1	1
67	0	1.960	0	0	1	1	0	0	1	1
68	0	1.927	0	0	1	1	1	0	0	1
69	0	1.832	0	0	1	1	1	0	0	1
70	0	1.941	0	1	1	1	0	0	1	1
71	1	0.099	1	0	1	1	1	0	0	1
72	0	1.714	0	0	1	1	0	0	0	1

73	1	0.151	0	1	1	1	1	0	0	1
74	0	1.697	0	0	1	1	1	0	0	1
75	0	1.692	0	1	1	1	1	1	0	1
76	0	0.214	1	1	0	1	1	0	0	1
77	0	1.624	0	0	1	1	1	1	0	1
78	0	1.566	1	0	1	1	1	1	0	1
79	0	1.528	0	0	1	1	1	0	0	1
80	0	1.520	0	0	1	1	1	0	0	1
81	1	0.487	1	1	1	1	1	0	1	0
82	0	1.481	0	0	1	1	1	1	0	1
83	1	1.410	0	1	1	1	1	0	0	1
84	0	0.003	0	0	1	1	1	0	0	1
85	0	1.259	0	1	1	1	1	1	0	1
86	0	1.205	0	0	1	0	1	0	0	1
87	0	1.180	0	0	1	1	1	0	0	1
88	0	0.572	1	0	0	0	1	1	0	1
89	0	1.120	0	1	1	1	1	0	0	1
90	0	1.103	0	0	1	1	1	0	0	1
91	0	1.065	0	1	1	1	1	0	0	1
92	0	0.498	1	0	0	0	1	1	0	1
93	0	0.991	0	0	1	1	1	0	0	1
94	0	0.991	0	0	1	1	1	0	0	1
95	0	0.994	0	0	1	1	1	0	0	1
96	1	0.898	0	0	1	1	1	0	0	1
97	0	0.969	0	0	1	1	1	0	0	1
98	0	0.895	0	0	1	1	1	0	0	1
99	1	0.893	0	0	1	1	1	0	0	1
100	0	0.701	0	0	0	0	1	0	0	1
101	0	0.810	0	0	1	1	1	0	0	1
102	0	0.742	0	1	1	1	1	0	0	1
103	0	0.758	0	0	1	1	1	0	0	1