UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Introducción al análisis de supervivencia para el curso de estadística III

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

MATEMÁTICO

P R E S E N T A :

COSSIO LORA GIANNI ATANASSIO

DIRECTOR DE TESIS:
M. en C. MARÍA DEL PILAR ALONSO REYES
2010

Agradezco al sudor de mi padre y a la fortaleza de mi madre.

Agradezco a mi tutora la maestra Ma. Del Pilar por su apoyo y paciencia para la realización del presente trabajo.

Agradezco a la maestra Lety por su apoyo, ya que gracias a ella se me facilito el acceso al material para el desarrollo del presente trabajo.

Índice general

Introducción **1**

1. Conceptos básicos **5**

1.1. Datos 5

1.2. Censura 6

2. Limitaciones **11**

2.1. Funciones de riesgo (Hazard) 11

2.1.1. Caso continuo 11

2.1.2. Caso discreto 14

2.2. Modelos paramétricos 18

2.2.1. Modelo exponencial 18

2.2.2. Modelo Weibull 19

2.2.3. Modelo gamma 20

2.2.4. Modelo log-normal 21

3. Tablas de Vida **23**

3.1. Estructura de una tabla de vida 23

3.2. Fórmula de Greenwood 25

3.3. Estimador Kaplan-Meier 30

3.3.1. Método de máxima verosimilitud 30

3.3.2. El estimador producto-límite 31

3.4. Estimación de la función de riesgo 36

3.4.1. Estimación de la función de riesgo acumulada 37

3.5. Prueba para dos grupos con datos de supervivencia 40
3.5.1. Prueba de logrank 41

4. Modelo de riesgo proporcional **45**
4.1. Modelo de regresión logística 45
4.2. Método de Newton-Raphson 48
4.3. Modelo de regresión de Cox 49
4.4. Estimación de los coeficientes de regresión del modelo de Cox 51
4.5. Intervalos de confianza y prueba de hipótesis 54
4.6. Estimación de la función de supervivencia y de riesgo base 55
4.7. Estratificación del modelo de Cox 59
4.7.1. Variables dependientes del tiempo 59
4.8. Residuales 60
4.8.1. Residual Cox-Snell 61
4.8.2. Residual martingala 63
4.8.3. Residual de desviación 63
4.8.4. Residual de Schoenfeld 64
4.8.5. Residual puntual (Score) 64

5. Ejemplo 1. Comparación de dos tratamiento de cáncer de próstata **67**

6. Ejemplo 2. Muestra de cáncer de ovario **75**

7. Conclusión **83**

A. Base de datos para muestra de cáncer de ovario **85**

Bibliografía **100**

Introducción

El desarrollo que compete a este trabajo de investigación hace referencia a un área de estudio de la estadística la cual tiene un alto índice de aplicación en diversos sectores del desarrollo científico del hombre, a está se le conoce como "Análisis de supervivencia"; en particular este trabajo se enfoca al tema de lo que es la teoría de supervivencia, y que se puede definir como la exploración de aquellos datos que son medidos en tiempo, de aquí que principalmente se tenga que la variable de respuesta y de observación este definida sobre intervalos de tiempo, el cual se puede escalar según las circunstancias de lo que se desea estudiar.

Las características principales que presentan los problemas que se buscan resolver a través del análisis de supervivencia es que cuando se realiza un muestreo de datos, en la mayoría de las veces estos no son continuos, ni aleatorios, más aun, los datos que se obtienen de los elementos que presentan o no el objeto de interés en el estudio son registrados de acuerdo a lapsos de tiempo en que tardó en ocurrir dicho objeto de interés; en la práctica ésta suele ser la falla en una unidad física (electrodomésticos, etc.) o la muerte si es biológico. Otra característica principal que presentan estos problemas, es que los elementos del estudio no están todos presentes desde el inicio de esté y tampoco los que entran se conservan hasta el final de tal suerte que se presentan en lapsos de tiempo de manera independiente.

Una muestra de lo que hace el análisis de supervivencia en la práctica puede ser observado primordialmente en las tablas de mortalidad, inclusive se refiere que es por la construcción de éstas que surgió la necesidad de herramientas estadísticas que las estudian, y por eso son conocidas estas técnicas como el análisis de supervivencia. Aunque el desarrollo de esta teoría no sólo tiene aplicación en mortalidad, sino que también puede ser empleada en diversos sectores. La presente investigación se realizó por el interés de conocer un área distinta de la estadística paramétrica y saber cómo poder resolver un problema el cual no puede ser representado, no necesariamente, por un modelo que se exprese. Además de que se busca motivar al lector para que conozca un enfoque distinto del que se usa por lo general en la estadística clásica de la manera de cómo se podría resolver un problema cuyos datos están determinados dentro de la variable tiempo. Otro interés general es que el trabajo de cierta forma sirva como guía introductora para aquellas personas que deseen desenvolverse y profundizar en el desarrollo de lo que es la teoría del análisis supervivencia. Mientras que lo que concierne al ámbito laboral es poder contar con herramientas alternas

para enfrentar y saber cómo solucionar un problema con datos que involucran tiempos de falla haciendo uso de la teoría descrita en esta sencilla investigación.

La manera con que se desarrollo éste trabajo es partiendo de la revisión de investigaciones antecesores que se enfocan principalmente a lo que es el análisis de supervivencia; como es el caso del libro citado en la bibliografía del autor David Collet, el cual está exclusivamente enfocado en el análisis de supervivencia con orientación en la investigación médica y otros libros que están inmersos en la teoría de la supervivencia como es el caso del libro "Survival Analysis"de Rupert G. Miller. También se revisaron otros libros los cuales quizá la teoría del análisis de supervivencia no eran sus tópicos principales, pero sí le dan cierta importancia. Una vez hecha esta revisión la estrategia con que se realizo esta basada en dar un orden que permitiera al lector ir entendiendo una conceptualización de cada capítulo descrito en este trabajo, sin saltarse o dar por entendido elementos, conceptos, definiciones, teoremas, etc.; esto de igual forma con el fin de que el interesado en este material no pierda la secuencia de los temas al encontrarse con cosas que puedan ser desconocidas o quizá no tenga presente en ese momento.

La estructura en este trabajo tiene una perspectiva general de los puntos o tópicos más importantes de lo que es en si el análisis de supervivencia. Una vez desarrollada la teoría correspondiente, se buscaron ejemplos sencillos los cuales ejemplifican la aplicación de la teoría haciendo uso en cierto momento de paquetes estadísticos, como Statistica, S-plus, etc.

Algo que se debe rescatar en este trabajo es que está diseñando para lectores de nivel superior que tengan una orientación en estadística aplicada o afín, y que posean un conocimiento quizá básico de probabilidad y estadística, es decir, que tengan presente conceptos de como se define la probabilidad de un evento, en probabilidad condicional, contraste de hipótesis, estimación entre otros para entender aquellos conceptos que son parte del proceso de investigación del análisis de supervivencia. Aunque también por la manera en que se estructura la presente tesis, puede servir de motivación para personas que no tengan un contacto directo con matemáticas aplicadas.

La distribución en que se presenta en este trabajo está dada en cinco capítulos los cuales desglosan paso a paso la información primordial del tema. En el capítulo uno se describen principalmente los conceptos fundamentales, es decir, se van a definir a aquellos sobre los cuales la teoría de supervivencia se basa. Estos son por ejemplo el tipo de datos que maneja el análisis de supervivencia, la falla, el evento, el dato de supervivencia, variable aleatoria, tiempo, etc.. En este capítulo también se describe la base más importante de la teoría de supervivencia, la "censura", mediante una definición general y luego se desglosará en diferentes tipos.

En el capítulo dos se verá con detalle cómo se define una variable aleatoria de supervivencia, y de igual forma una función de riesgo. Hay que notar que muchos libros o otros estudios, como en la mortalidad, esta función es conocida como "hazard". Las funciones referidas anteriormente tienen como plataforma una función de densidad de probabilidad definida especialmente para datos de supervivencia. También hay que notar que estas tres funciones se describen en dos casos, continuo y discreto, y que están mutuamente

relacionadas. En este capítulo se hará notar los diferentes tipos de modelos de riesgo o supervivencia paramétricos, llamados así porque en ciertos tipos de problemas pueden ser expresados por una función de densidad determinada por uno o varios parámetros y a veces esta se reconoce fácilmente; algunos de los modelos paramétricos más importantes son el modelo exponencial, Weibull, Gamma entre otros,etc..

En el capítulo tres se analiza una aplicación importante de la teoría de supervivencia, las tablas de vida. Se describirá los elementos que se usan para su construcción y los pasos que se deben hacer para obtenerla, se verá el papel que juega la función de supervivencia dentro de dicha tabla, y se concluirá dando una expresión de cómo debe estar estructurada. En el transcurso de la construcción de la tabla de vidá será necesario estimar la función de supervivencia y con ella el estimador conocido como actuarial. También en este capítulo se obtendrá el estimador kaplan-Meier o producto para la función de supervivencia. Una vez realizada la estimación para la función de supervivencia de la relación obtenida en el capitulo dos se prosigue a estimar la función de riesgo y su acumulada. Una vez hecho esto se verá otra aplicación, la cual involucra probar si dos grupos de estudio pueden ser descritos por una sola función de supervivencia, esta prueba es conocida como "logrank".

En el capítulo cuatro se observará uno de los aspectos más importantes en el análisis de supervivencia. Se buscara analizar y describir de la manera más sencilla lo que es el modelo de riesgo proporcional, el cual se representa por parámetros regresores correspondientes sus variables independientes. De forma muy básica se verá lo que es un modelo de regresión logística y el método de Newton-Rapson, el cual será de gran utilidad para poder estimar los parámetros. Después se describirá el modelo de riesgo proporcional, conocido en la literatura como regresión de Cox, también será estimado y se le calcularán los intervalos de confianza. Por último se dará una pequeña introducción a residuales que se usan en el modelo de Cox, notando que éstos sólo son descritos como motivación y no como una parte fundamental de este trabajo.

En el capítulo cinco se darán unos ejemplos que permitan ver como se trabaja con la teoría desarrollada a lo largo de la tesis. En algunos de estos ejemplos se harán cálculos hechos "a pie"para hacer notar que la teoría funciona, pero en la mayoría de los ejemplos, por el hecho de que la base de datos es muy grande, se hizo uso de paquetes estadísticos los cuales se refirieron en un principio para obtener el resultado buscado en cada ejemplo, además de dar un análisis interpretativo en cada ejemplo.

Por último se expresan las conclusiones y los anexos correspondientes.

Capítulo 1

Conceptos básicos

El análisis de supervivencia comprende un conjunto de procedimientos estadísticos que involucran el estudio de distribuciones del tiempo de vida puesto que la variable de respuesta de interés es *el tiempo hasta la ocurrencia de un evento* y por el cual entiende por vida de personas, robots, software, etc., es decir, todo aquello que comprenda un deterioro de funciones, ya sea en lo industrial o lo biológico.

1.1. Datos

Describir los datos de una cierta muestra en la forma del tiempo es la base del análisis de supervivencia puesto que éste se encarga de analizarlos y además que su variable de respuesta es el tiempo medido entre dos sucesos. Se acostumbra que la variable de respuesta mida el transcurso entre el inicio de un suceso (tiempo de origen) y la consecuencia de un segundo suceso. Este ultimo en ciertos casos suele ser el esperado por el investigador (en esta situación dicho evento será llamado punto final, **evento** o **falla**) o ajeno al estudio, cuando es de la segunda forma se dice que este suceso da información censurada, este punto será tratado más adelante.

En investigación médica, el tiempo de origen corresponde al ingresar un individuo a un estudio (hay que notar que no todos los individuos ingresan al mismo tiempo). Si el punto final de un individuo es la muerte, el dato que resulta es tiempo de supervivencia (**dato de supervivencia**). Sin embargo si dicho punto no ocurre se puede rescatar información del individuo, la cual se dice que está censurada.

Típicamente el valor de la variable aleatoria es el tiempo para la falla de un componente, ya sea de lo industrial o biológico.

Por **tiempo** se entenderá días, semanas, meses o años desde el punto inicial (tiempo de origen) de un componente del estudio hasta que ocurre un evento. Por ejemplo la edad de un individuo hasta que ocurre o no el evento, la vida funcional.

Por **evento** se entenderá la muerte, una enfermedad incidente, tratamientos médicos o cualquier designación experimental que pueda ser de ayuda a los componentes del estudio.

Dato de supervivencia, es el término usado para describir medida del tiempo de algún evento.

En el desarrollo de esté trabajo la variable aleatoria comúnmente se le conoce como **tiempo de supervivencia**. Debido a que el individuo en estudio sobrevivió cierto tiempo en el estudio y se hace referencia si se presento o no el evento. Para determinar un tiempo de falla se necesitan de tres condiciones principalmente:

i. Un tiempo de origen que debe ser definido inequívocamente (tiempo calendario).

ii. Definir una escala para medir el tiempo transcurrido durante el estudio.

iii. Tener bien definido el significado de falla. Para que quede enteramente clara en el estudio cuando se presente.

El tiempo calendario debería ser definido con precisión para cada individuo, el cual no es necesariamente el mismo para todos, al igual que el tiempo de falla. Una escala para medir la supervivencia es el tiempo reloj,o real, ya que permite precisar la falla.

1.2. Censura

El hecho de que el tiempo se mida secuencialmente tiene como consecuencia la censura, la cual se presenta cuando se tiene información incompleta o si los tiempos de supervivencia no se conocen con exactitud, a los datos con estas características se les conoce como censurados. Los datos del estudio pueden estar sesgados debido a las pérdidas de seguimiento, el fin del estudio o a la censura.

En el análisis de supervivencia se asume un supuesto básico: los mecanismos del evento esperado en el estudio y la censura son estadísticamente independientes, o el sujeto censurado es representativo de los que sobreviven, es decir, los no censurados representan bien a los que si lo están.

La terminación de la observación puede ser controlada de distintas maneras. Las más usuales son efectuadas por los siguientes dos esquemas, respectivamente:

I. La terminación de la observación se da en un punto del tiempo predefinido. El dato resultante es truncado.

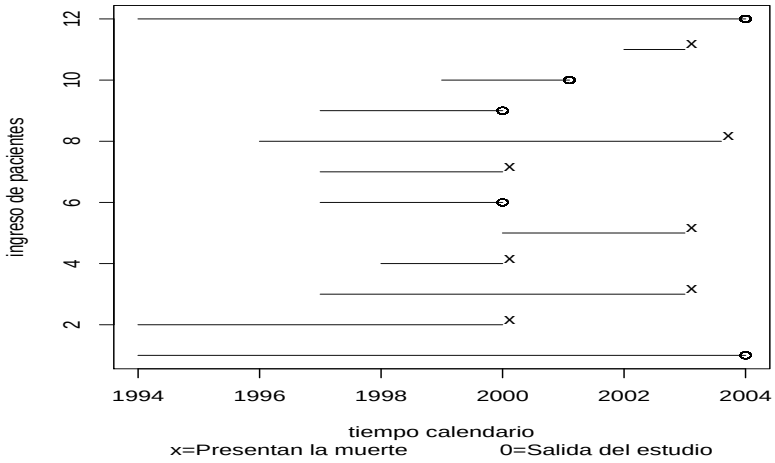
II. La terminación de la observación se da cuando un número predefinido de eventos ha ocurrido. El dato resultante es llamado censurado.

Algunos autores llaman al esquema I y II Censura tipo I y II respectivamente. Es importante que en los esquemas, el tiempo de finalización y el número de fallas, sean variables aleatorias.

Censura: también se puede ver como la técnica para el tiempo de estudio reducido. En el análisis de supervivencia las observaciones pueden ser de una duración indefinidamente larga. Es por esto que estos son a tiempos de niveles manejables.

En el siguiente ejemplo gráfico se muestra como los datos con los que se trabaja en análisis de supervivencia son aleatorios, así también como se presenta datos censurados y no censurados, estos últimos son los que alcanzan el evento esperado.

Ejemplo 1. *Pacientes con cáncer terminal. En un hospital se lleva a cabo un estudio con 12 enfermos, el cual tendrá una duración de 11 años. La incorporación de los enfermos no se produce al mismo tiempo, por lo que el registro de tiempo será distinto. El evento esperado (la muerte), será medido desde el ingreso (punto de inicio) de cada enfermo. Véanse las gráficas (1.1) y (1.2).*



Gráfica 1.1: Datos en tiempo calendario

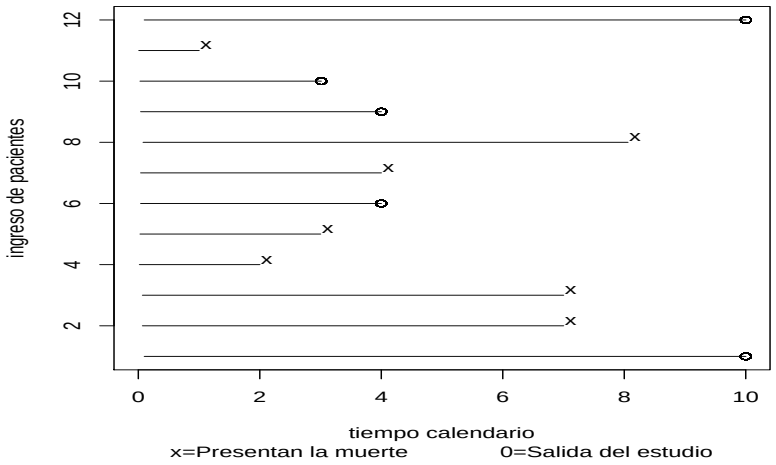
Tipos de censura: Una observación censurada contiene información parcial acerca de las variables aleatorias que son de interés para el estudio.

Censura tipo I

Existen situaciones donde el tiempo o el ingreso no son suficientes para realizar el estudio, por lo que el investigador debe de determinar un límite para esté ultimo, cuando se presenta una duración predefinida del estudio a esto se le conoce como *tiempo de censura fijo*. Esté tipo de censura se puede escribir matemáticamente.

Sean $T_1, T_2, \cdots, T_n$ independientes e idénticamente distribuidas (i.i.d.) con función de densidad (f.d.) F.

Sea t_c algún número fijo (predesignado) al que se llama *tiempo de censura fijo*. En vez de tener que observar $T_1, T_2, \cdots, T_n$ (la variable aleatoria de interés) se puede observar



Gráfica 1.2: Datos en tiempo en el estudio

solamente $Y_1, Y_2, \cdots, Y_n$ donde

$$Y_i = \begin{cases} T_i & \text{si} \quad T_i \leq t_c, \\ t_c & \text{si} \quad t_c < T_i \end{cases}$$

Nótese que la función de distribución de Y tiene un sentido positivo, $P[T > t_c] > 0$ si $Y = t_c$.

Censura tipo II

Este tipo de censuras se da debido a que el investigador observa solamente las primeras r fallas (eventos) de n posibles ($r < n$) como consecuencia de ciertas razones, como el tiempo, el conocimiento empírico del investigador, los ingresos, etc. y decide tomar el valor de la última falla observada para el resto de las observaciones. Escrito matemáticamente.

Sea $r < n$ fija, y sean $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(n)}$ las estadísticas de orden de $T_1, T_2, \cdots, T_n$. Se realizan observaciones y se detiene hasta que ocurre la r-ésima falla. Por lo que se tiene $T_{(1)}, T_{(2)}, \cdots, T_{(n)}$. La muestra ordenada es:

$$\begin{array}{ccc} Y_{(1)} & = & T_{(1)} \\ Y_{(2)} & = & T_{(2)} \\ \vdots & \vdots & \vdots \\ Y_{(r)} & = & T_{(r)} \\ Y_{(r+1)} & = & T_{(r)} \\ \vdots & \vdots & \vdots \\ Y_{(n)} & = & T_{(r)} \end{array}$$

Se observa hasta la r-ésima falla y luego se asigna el valor $T_{(r)}$ al resto de las variables aleatorias Y_i para $i > r$.

Las censuras del tipo I y II tienen sus mayores aplicaciones en los sectores industriales.

Censura aleatoria

La censura aleatoria tiene su mayor aplicación en el sector biológico. En este caso el investigador no tiene control sobre el tiempo censurado. Existen generalmente 3 razones por las cuales ocurre este tipo de situaciones.

1. Un individuo del estudio no experimenta la falla antes del final del estudio.
2. Un individuo del estudio es retirado de éste. Es decir, no se sabe si experimenta la falla o no.
3. Un individuo del estudio experimenta la falla por razones externas al estudio.

Sean $C_1, C_2, \cdots, C_n$ i.i.d. con f.d. G. C_i es el tiempo de censura asociado con T_i, correspondiente al i-ésimo individuo.

Solamente se puede observar $(Y_1, \delta_1), \cdots, (Y_n, \delta_n)$ donde:

$Y_i = min(T_i, C_i)$, y

$$\delta_i = \begin{cases} 1 \text{ si } T_i \leq C_i, \text{ si, } T_i \text{ no es censurado} \\ 0 \text{ si } T_i > C_i, \text{si, } T_i \text{ es censurado} \end{cases}$$

Nótese que $Y_1, Y_2, \cdots, Y_n$ son i.i.d. con alguna f.d. H. Así mismo $\delta_1, \cdots, \delta_n$ contiene información censurada.

El suponer que los tiempos de falla y de censura son independientes parece justificado ya que se tienen ingresos y ocurren pérdidas, ambos de secuencia aleatoria.

Censura aleatoria por la derecha y la izquierda

Censura por la derecha. Una observación se dice que es censurada por la derecha si el tiempo de falla T se desconoce y sólo se tiene el dato de que es mayor que el tiempo en que se registra la censura. De igual manera, se dice que una observación es *censurada por la izquierda* si se tiene el dato de que el valor desconocido del tiempo de falla T es menor que el tiempo de censura.

Para la censura por la izquierda se puede observar solamente $(Y_1, E_1), \cdots, (Y_n, E_n)$ donde $Y_i = max(T_i, C_i)$, y

$$E_i = \begin{cases} 1 \text{ si } C_i \leq T_i, \text{ si, } T_i \text{ no es censurado} \\ 0 \text{ si } C_i > T_i, \text{si, } T_i \text{ es censurado} \end{cases}$$

Tanto la censura por la izquierda como por la derecha son casos particulares de la *Censura por intervalo*, de tal forma que se puede observar que la variable aleatoria de

interés cae dentro de un intervalo. Si T_i es censurada por la derecha y la izquierda entonces se puede observar que T_i cae en el intervalo $[C_i, \infty)$ y $[0, C_i]$ respectivamente.

Ejemplo 2. *De censura por la izquierda y la derecha.*
Un psicólogo desea estudiar el tiempo en el que un cierto grupo de 10 niños de 11 años tarda en aprender a ejecutar una habilidad en 30 días. Cuando comenzó el estudio 3 niños ya ejecutaba dicha habilidad, estos representan la censura por la izquierda, otros 4 aprendieron a ejecutar la habilidad en 3, 15, 22 y 23 días, estos niños representan las observaciones no censuradas, mientras que los 3 restantes que no aprendieron a ejecutar la habilidad en los 30 días, representan la censura por la derecha.

Capítulo 2

Limitaciones

Se debe de tener en cuenta que existen diversos motivos por los que los datos de supervivencia no se manejan con los métodos estadísticos usuales. Una de las razones principales se debe a que los estos no siguen el comportamiento de una distribución simétrica. "Un histograma construido de los tiempos de supervivencia de un grupo de individuos similares tiende a tener una cola positiva, es decir, tendrá una larga cola a la derecha del intervalo que contiene el número más grande de observaciones. Como consecuencia, esto no será una razón para asumir que los datos siguen el comportamiento de una distribución normal". Otro motivo es que la variable de respuesta está dada en tiempo por lo que no se puede medir como otras variables, es decir, que los datos en análisis de supervivencia son una medida de tiempo a una respuesta, fallo, muerte, recaída o algún evento esperado.

2.1. Funciones de riesgo (Hazard)

En el análisis de datos de supervivencia se usan tres principales funciones, las cuales permiten comprender el comportamiento de la variable aleatoria T (tiempo en que se espera el evento), y son:

- Función de supervivencia, $s(t)$.
- Función de densidad de probabilidad (f.d.p.), $f(t)$.
- Función de riesgo o función de hazard, $h(t)$.

2.1.1. Caso continuo

Definición 1. *Definimos una* **variable aleatoria de supervivencia** *T, si t es un resultado observado de T que cae dentro del intervalo* $[0, \infty)$.

Sea $F(.)$ la **función de distribución acumulada** (f.d.a.) de T que corresponde a la **función de densidad de probabilidad** (f.d.p.) $f(.)$.
Note que $f(t) = 0\ \forall t < 0$, entonces

$$F(t) = P(T \leq t) = \int_0^t f(x)dx \tag{2.1}$$

La probabilidad de que un elemento de una muestra sobreviva más de un tiempo t está dada por **la función de supervivencia**

$$s(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(x)dx \tag{2.2}$$

Esta función también es llamada como *función de fiabilidad.*

Observación 1. *Por la densidad de f pasa lo siguiente:*
$s(0) = \int_0^\infty f(x)dx = 1$
$s(\infty) = \lim_{t\to\infty} \big(1 - F(t)\big) = 0$
$\therefore$ *$s(t)$ es monótona decreciente y continua por la izquierda.*

Se define la **función de densidad de probabilidad** (f.d.p.) como el límite de la razón entre la probabilidad de que un elemento de la muestra presente el evento en el intervalo $(t, t + \Delta t)$ y un incremento en t:

$$f(t) = \lim_{\Delta t\to 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t} \tag{2.3}$$

Se sabe que:

$$\begin{aligned} P(t < T \leq t + \Delta t) &= P(T \leq t + \Delta t) - P(T \leq t) \\ &= F(t + \Delta t) - F(t) \\ &= 1 - s(t + \Delta t) - (1 - s(t)) \\ &= s(t) - s(t + \Delta t) \\ &= -(s(t + \Delta t) - s(t)) \end{aligned}$$

$$\begin{aligned} \Rightarrow\ f(t) &= \lim_{\Delta t\to 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t} \\ &= \lim_{\Delta t\to 0} \frac{-(s(t + \Delta t) - s(t))}{\Delta t} \\ &= -\lim_{\Delta t\to 0} \frac{(s(t + \Delta t) - s(t))}{\Delta t} \\ &= -\frac{ds(t)}{dt} \end{aligned}$$

$$\therefore\ f(t) = -\frac{ds(t)}{dt} \tag{2.4}$$

La función de riesgo o función hazard se define como el límite de la razón de la probabilidad de que un elemento de la muestra a estudiar presente el evento en un intervalo pequeño $(t, t + \Delta t)$, dado que sobrevivió al tiempo t y un incremento en t:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T \geq t)}{\Delta t} \tag{2.5}$$

por otro lado

$$\begin{aligned} P(t < T \leq t + \Delta t | T \geq t) &= \frac{P(t < T \leq t + \Delta t)}{P(T \geq t)} \\ \Rightarrow\ h(t) &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t (P(T \geq t))} \\ &= \frac{f(t)}{s(t)} \end{aligned}$$

$$\therefore\ h(t) = \frac{f(t)}{s(t)} \tag{2.6}$$

Observación 2.

$$\begin{aligned} h(t) &= \frac{f(t)}{s(t)} \\ &= \frac{-\frac{ds(t)}{dt}}{s(t)} \\ &= \frac{-d\log(s(t))}{dt} \end{aligned}$$

$$\therefore\ h(t) = \frac{-d\log(s(t))}{dt} \tag{2.7}$$

Si se integra la función de riesgo $h(.)$ en el intervalo $(0, t)$ se obtiene la *función de riesgo acumulada.*

$$H(t) = \int_0^t h(x)dx \tag{2.8}$$

$$\Rightarrow\ H(t) = \int_0^t \left(\frac{-d\log(s(x))}{dx}\right)dx$$

$$\begin{aligned} &\therefore\ H(t) = -\log(s(t)) \\ &\Rightarrow\ s(t) = \exp(-H(t)) \end{aligned} \tag{2.9}$$

$$\therefore\ s(t) = \exp\left(-\int_0^t h(x)dx \right) \tag{2.10}$$

y como $f(t) = -\dfrac{ds(t)}{dt}$

$$\therefore\ f(t) = h(t)\exp\left(-\int_0^t h(x)dx \right) \tag{2.11}$$

Nótese que las expresiones 2.7, 2.10 y 2.11 están mutuamente relacionadas.

2.1.2. Caso discreto

Cuando los datos de supervivencia representan un número contable de ciclos de alguna clasificación. En este caso se dice que T es una variable aleatoria discreta.

Sea T una variable aleatoria discreta que toma valores $0 \leq t_1 \leq t_2 \leq \cdots$ entonces $f(t_j)$ f.d.p se define como:

$$f(t) = \begin{cases} P(T = t) & \text{si } t = t_j,\ j = 1, 2, \cdots \\ 0 & \text{si } t \neq t_j \end{cases} \tag{2.12}$$

La función de supervivencia está dada por:

$$s(t) = P(T \geq t) = \sum_{j|t_j \geq t} f(t_j) \tag{2.13}$$

Observación 3. *Dado que $f(t)$ es función de densidad, $s(t)$ satisface:*
$s(0) = P(T \geq 0) = \sum_{j|t_j \geq 0} f(t_j) = 1$
$s(\infty) = P(T \geq \infty) = 0$
$\therefore$ *$s(t)$ es monótona decreciente y continua por la izquierda.*

La función de riesgo (hazard) esta definida como la probabilidad de ocurrir el evento esperado en $t = t_j$ dado que sobrevivió antes de t_j:

$$h(t_j) = P(T = t_j | T \geq t_{j-1}) \tag{2.14}$$

Como $P(T = t_j | T \geq t_{j_1}) = \dfrac{P(t_{j-1} \leq T = t_j)}{P(t \geq t_{j-1})}$

$$\Rightarrow\ h(t_j) = \frac{f(t_j)}{s(t_j)} \tag{2.15}$$

Observación 4. $f(t_j) = s(t_j) - s(t_{j+1})$

$$\begin{aligned}
\text{pues } s(t_j) - s(t_{j+1}) &= \sum_{k|t_k \geq t_j} f(t_k) - \sum_{k|t_k \geq t_{j+1}} f(t_k), \quad j = 1, 2, \ldots \\
&= \sum_{k|t_k \geq t_j} P(T = t_k) - \sum_{k|t_k \geq t_{j+1}} P(T = t_k), \quad j = 1, 2, \ldots \\
&= P(T = t_j)
\end{aligned}$$

$$\text{Como } h(t_j) = \frac{f(t_j)}{s(t_j)} \;\Rightarrow\; h(t_j) = \frac{s(t_j) - s(t_{j+1})}{s(t_j)}$$

$$h(t_j) = 1 - \frac{s(t_{j+1})}{s(t_j)} \tag{2.16}$$

Despejando $s(t_{j+1})$ de 2.16 se tiene que

$$s(t_{j+1}) = (1 - h(t_j))(s(t_j))$$

Recursivamente se tiene que

$$\begin{aligned}
s(t_j) &= (1 - h(t_{j-1}))s(t_{j-1}) \\
s(t_{j-1}) &= (1 - h(t_{j-2}))s(t_{j-2}) \\
&\vdots \quad \vdots \quad \vdots \\
s(t_2) &= (1 - h(t_1))s(t_1) \\
s(t_1) &= 1
\end{aligned}$$

$$\Rightarrow\; s(t_j) = (1 - h(t_{j-1}))(1 - h(t_{j-2})) \cdots (1 - h(t_2))(1 - h(t_1))$$

por lo que la función de supervivencia puede ser también dada por

$$s(t_j) = \prod_{k|t_k < t_j} (1 - h(t_k)) \tag{2.17}$$

$$\text{Dado que } h(t_j) = \frac{f(t_j)}{s(t_j)} \;\Rightarrow\; f(t_j) = h(t_j)s(t_j)$$

$$f(t_j) = h(t_j) \prod_{k|k<j} (1 - h(t_j)) \tag{2.18}$$

Lo que determina una relación mutua de las funciones de densidad, de supervivencia y de riesgo.

Función de supervivencia empírica

Es muy usual que si en el estudio no se encuentran datos censurados la función de supervivencia se estime como el número de elementos de la muestra que sobrevivieron más allá del tiempo t y se define como:

$$\hat{s}(t) = \frac{\#\text{ de elementos que sobrevivieron más del tiempo } t}{\text{total de elementos de la muestra}} \tag{2.19}$$

A esta función también se le conoce como la supervivencia empírica.

Al igual que la función de supervivencia, si no existen datos censurados, la f.d.p f se estima como el número de elementos de la muestra que presentan el evento esperado en un intervalo de tiempo y se define como:

$$\hat{f}(t) = \frac{\#\text{ de elementos que presentaron el evento en el intervalo } (t, t+\Delta t)}{(\text{total de elementos de la muestra})(\text{longitud del intervalo})} \tag{2.20}$$

Observación 5. *La curva de densidad, gráfica de $f(t)$, da una muestra de que al comienzo del estudio el rango de fallo (o que se presente el evento) es grande y éste decrece cuando el tiempo se incrementa.*

La función de densidad de probabilidad estimada es conocida también como el rango de falla incondicional.

De igual forma si no se presentan datos censurados en el estudio la función de riesgo se estima como el número de elementos de la muestra que presentan el evento esperado en un intervalo de tiempo; dado que ellos sobrevivieron en el comienzo del intervalo se define como:

$$\hat{h}(t) = \frac{\#\text{ de elementos que presentaron el evento en el intervalo } (t, t+\Delta t)|(T > t)}{(\text{total de sobrevivientes en t})(\text{longitud del intervalo})} \tag{2.21}$$

La función de riego estimada es conocida también como el rango de falla instantáneo, fuerza de mortalidad.

Ejemplo 3. *Las tres primeras columnas de la tabla 2.1 muestran los datos de 40 pacientes con Myeloma. Los tiempos de supervivencia son agrupados en tiempos de 5 meses.*

Se muestra sólo para el intervalo que va de 10 hasta 15 meses para ejemplificar el calculo referido en la tabla (2.1):

$$\hat{s}(10) = \frac{28}{40} = 0{,}700 \quad \hat{f}(10) = \frac{6}{40 * 5} = 0{,}030 \quad \hat{h}(10) = \frac{6}{28 * 5} = 0{,}042$$

Otra manera de ver matemáticamente la función de supervivencia estimada es que dadas $T_1, \ldots, T_n$ variables aleatorias de supervivencia i.i.d se define $\forall t$

$$s_n(t) = \frac{1}{n} \sum_{i=1}^{n} I_{(t,\infty)}(T_i) \tag{2.22}$$

donde $I_{(t,\infty)}$ es la función indicadora de un evento.

Tiempo de supervivencia t (meses)	# de pacientes que sobrevivieron al comienzo del intervalo	# pacientes que murieron en el intervalo	$\hat{s}(t)$	$\hat{f}(t)$	$\hat{h}(t)$
0-5	40	5	1.000	0.025	0.025
5-10	35	7	0.875	0.035	0.040
10-15	28	6	0.700	0.030	0.042
15-20	22	4	0.550	0.020	0.036
20-25	18	5	0.450	0.025	0.055
25-30	13	4	0.325	0.020	0.061
30-35	9	4	0.225	0.020	0.088
35-40	5	0	0.125	0.000	0.000
40-45	5	1	0.125	0.005	0.040
≥ 50	4	2	0.100	0.010	0.100

Tabla 2.1: **Datos de supervivencia y estimación de las funciones de supervivencia, densidad y riesgo de 40 pacientes con Myeloma.**

Ejemplo 4. Función de supervivencia para el efecto placebo.

Gehan (1965), Lawless(1982) y otros autores han discutido datos de ensayos clínicos examinando tiempos (semanas) de remisión de esteroides inducidos para pacientes con Leucemia. Un grupo de 21 pacientes fue tratado con 6-mercatopurina (6-MP); un segundo grupo de 21 pacientes fue tratado con placebo. Los datos $t_1, \ldots, t_n$ estan dados en la tabla (2.2), y son el tiempo de remisión de este grupo de placebos.

Tiempos (semanas) de remisión de esteroides inducidos							
1	1	2	2	3	4	4	5
5	8	8	8	8	11	11	12
12	15	17	22	23			

Tabla 2.2: **Grupo placebo.**

Se mostrará como estimar la probabilidad de más de 3 meses de remisión para pacientes en el grupo placebo.

Se usa $s_{21}(12)$ para estimar $s(12)$, la probabilidad acertada de sobrevivir más de 12 semanas. Se encuentra que 4 pacientes vivieron más de 12 semanas, esto significa que 4 indicadores de la función $I_{(12,\infty)(t_i)}$, para $i = 1, \ldots, 21$ tienen valor de 1, los 17 restantes tienen valor de 0.

$$\therefore\ s_{21}(12) = \hat{s}(12) = \frac{1}{21}\sum_{i=1}^{21} I_{(12,\infty)}(t_i) = \frac{4}{21}$$

2.2. Modelos paramétricos

2.2.1. Modelo exponencial

La distribución exponencial tiene un papel importante en los estudios de supervivencia semejante al de la distribución normal en otras áreas de la estadística. Ésta tiene su mayor aplicación en el sector industrial, ya que no considera si se afecta las funciones de un elemento biológico en el futuro; la distribución exponencial describe datos con un mismo riesgo.

La distribución exponencial es representada por un sólo parámetro λ, el cual representa la constante de riesgo. Si el valor de λ es alto entonces el riesgo es alto y la supervivencia poca; si el valor de λ es bajo entonces el riego es mínimo y la supervivencia mayor. Si el valor λ es 1 la distribución es llamada *distribución exponencial unitaria.*

Si T es una variable aleatoria de supervivencia y el valor de riesgo es constante entonces T se distribuye exponencialmente con parámetro λ.

Sea $T \sim \exp(\lambda)$ entonces se define la función de riesgo como:

$$h(t) = \lambda,\ t \geq 0 \tag{2.23}$$

Por 2.10 se obtiene la función de supervivencia

$$\begin{aligned} s(t) &= exp\Big(-\int_0^t \lambda dx \Big) \\ &= exp\Big(-\lambda \int_0^t dx \Big) \\ &= exp(-\lambda t) \end{aligned}$$

$$\therefore\ s(t) = exp(-\lambda t) \tag{2.24}$$

y por 2.4 se obtiene la función de densidad de probabilidad:

$$f(t) = \lambda \exp(-\lambda t) \tag{2.25}$$

La media y la varianza del tiempo de supervivencia para el modelo exponencial están dadas por,

$$E(T) = \frac{1}{\lambda}$$

$$var(T) = \frac{1}{\lambda^2}$$

Si se busca extender la distribución exponencial a dos parámetros λ, G donde es segundo es un tiempo de garantía dentro del cual no puede ocurrir falla o muertes, o representa un tiempo de supervivencia mínimo.

Entonces la función de densidad de probabilidad es:

$$f(t) = \begin{cases} \lambda \exp(-\lambda(t-G)) & ; t \geq G \\ 0 & ; t < G \end{cases} \tag{2.26}$$

de 2.2 se tiene la función de supervivencia

$$s(t) = \begin{cases} \exp(-\lambda(t-G)) & ; t \geq G \\ 0 & ; t < G \end{cases} \tag{2.27}$$

y de 2.6 se tiene la función de riesgo

$$h(t) = \begin{cases} \lambda & ; t \geq G \\ 0 & ; t < G \end{cases} \tag{2.28}$$

y el valor medio de tiempo de supervivencia es $G + \dfrac{1}{\lambda}$.

Observación 6. *Si $G = 0$ se tiene la distribución exponencial normal.*

2.2.2. Modelo Weibull

La distribución Weibull se considera como una generalización de la distribución exponencial, pero diferente de ella. Ésta se ha usado para establecer estados de confiabilidad, la mortalidad de enfermedades humanas, así también como la vida útil de un elemento físico, está se determina por los parámetros α, β, el parámetro α, es llamado de forma, dado que establece como es la curva mientras que β, es llamado de escala, puesto que determina el tamaño de las unidades en que se mide la variable aleatoria.

La distribución de Weibull se usa cuando el riesgo se incrementa o decrece el tiempo en relación con la supervivencia.

La función de densidad de probabilidad es:

$$f(t) = \beta\alpha(\beta t)^{\alpha-1} \exp[-(\beta t)^{\alpha}] \tag{2.29}$$

de 2.2 se tiene la función de supervivencia

$$s(t) = \exp[-(\beta t)^{\alpha}] \tag{2.30}$$

y de 2.6 se tiene la función de riesgo

$$h(t) = \beta\alpha(\beta t)^{\alpha-1} \tag{2.31}$$

con media y variación de supervivencia $E(T)$ y $var(T) = E(T^2) - E^2(T)$ respectivamente. Dadas por el $r-$ésimo momento $E(T^r)$

$$\begin{aligned} E(T^{\,r}) &= \int_o^\infty t^r \beta\alpha(\beta t)^{\alpha-1} \exp[-(\beta t)^\alpha] dt \\ &= \alpha\beta^{1-r} \int_o^\infty (\beta)^{\alpha+r-1} \exp[-(\beta t)^\alpha] dt \\ \text{si } w = (\beta t)^\alpha &\Rightarrow t = \beta^{-1} w^{\frac{1}{\alpha}} \\ &\Rightarrow dt = \beta^{-1}\alpha^{-1} w^{\frac{1}{\alpha}-1} dw \\ \Rightarrow E(T^{\,r}) &= \alpha\beta^{1-r}\alpha^{-1}\beta^{\alpha+r-1}\beta^{-\alpha-r+1}\beta^{-1} \int_o^\infty w^{1+\frac{r}{\alpha}-\frac{1}{\alpha}} w^{\frac{1}{\alpha}-1} dw \\ &= \beta^{-r} \int_o^\infty w^{\frac{r}{\alpha}} \exp(-w) dw \\ &= \beta^{-r}\Gamma(1+\frac{r}{\alpha}) \end{aligned}$$

$$\therefore\ E(T) = \frac{\Gamma(1+\alpha^{-1})}{\beta}$$

$$\therefore\ var(T) = \frac{1}{\beta^2}\left[\Gamma(1+\frac{2}{\alpha}) - \Gamma^2(1+\frac{1}{\alpha})\right]$$

donde $\Gamma(\delta)$ es la función gamma

$$\begin{aligned} \Gamma(\delta) &= \int_0^\infty x^{\delta-1} e^{-x} dx \\ &= (\delta-1)! \text{ donde } \delta \text{ es un valor entero.} \end{aligned}$$

Si de igual forma que en el caso exponencial, se agrega un parámetro G a la distribución de Weibull, en donde el parametro agregado es un tiempo de garantía de que no ocurra una falla. Entonces se tiene que:

$$f(t) = \beta^\alpha (t-G)^{\alpha-1} \exp[-\beta^\alpha (t-G)^\alpha]$$

$$s(t) = \exp[-\beta^\alpha (t-G)^\alpha]$$

$$h(t) = \beta^\alpha (t-G)^{\alpha-1}$$

2.2.3. Modelo gamma

La función de distribución gamma ha sido usada comúnmente como modelo para establecer los problemas de confiabilidad de productos de fábricas. Está determinada por dos parámetros α y β.

Cuando $0 < \beta < 1$ " se dice que hay deterioro negativo y el valor de riesgo decrece monótonamente desde ∞ hasta β si el tiempo de el estudio se incrementa desde 0 hasta ∞. Cuando $\alpha > 1$ se dice que hay deterioro positivo y el valor de riesgo se incrementa monótonamente desde 0 hasta 1 si el tiempo se incrementa desde cero hasta ∞. Cuando $\alpha = 1$ se tiene el caso exponencial". Para la distribución gamma se tiene la f.d.p.:

$$f(t) = \frac{\beta(\beta t)^{\alpha-1}\exp[-\beta t]}{\Gamma(\alpha)} \ ; \ t > 0, \ \alpha, \ \beta > 0 \tag{2.32}$$

donde $\Gamma(\alpha)$ es la ya referida función gamma. De 2.2 se tiene la función de supervivencia

$$s(t) = \int_t^\infty \frac{\beta(\beta u)^{\alpha-1}\exp[-\beta u]du}{\Gamma(\alpha)} \tag{2.33}$$

y de 2.6 se tiene la función de riesgo

$$h(t) = \frac{\beta(\beta u)^{\alpha-1}\exp[-\beta u]}{\displaystyle\int_t^\infty \beta(\beta u)^{\alpha-1}\exp[-\beta u]du} \tag{2.34}$$

Con tiempo de media y varianza de supervivencia:

$$E(T) = \frac{\alpha}{\beta} \text{ y } var(T) = \frac{\alpha}{\beta^2}.$$

2.2.4. Modelo log-normal

Al igual que la distribución Weibull la log-normal ha sido utilizada como modelo de distribución de vida, falla en sistemas eléctricos y la aparición de cáncer en pulmones entre otras aplicaciones. El uso de la distribución log-normal se da por el hecho de que los valores acumulados de Y, una variable aleatoria de supervivencia i.i.d., se pueden obtener de las tablas de la distribución normal estándar con media μ y varianza σ^2 y lo que se hace es que se toma $Y = \log T$ por lo que los valores de t pueden ser obtenidos de los antilogaritmos.

La función de densidad de probabilidad de Y está dada por:

$$f^*(y) = \frac{1}{(2\pi)^{1/2}\sigma}\exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right] \ ; \ -\infty < y < \infty$$

ahora si se toma $Y = \log T$ evaluado en una observación $T = t$ entonces se obtiene la función de densidad de probabilidad de T:

$$f(t) = \frac{1}{(2\pi)^{1/2}\sigma t}\exp\left[-\frac{1}{2}\left(\frac{\log t-\mu}{\sigma}\right)^2\right] \ ; \ t > 0 \tag{2.35}$$

Si se retoma (2.2) entonces se tiene que la función de supervivencia para el modelo log-normal es:

$$s(t) = 1 - \frac{1}{(2\pi)^{1/2}\sigma}\int_0^t \frac{1}{u}\exp\left[-\frac{1}{2}\left(\frac{\log u-\mu}{\sigma}\right)^2\right]du \tag{2.36}$$

$$\text{si } a = exp[-\mu] \;\Rightarrow\; -\mu = \log a$$

$$\Rightarrow\; \log u - \mu = \log u + \log a = \log(au)$$

$$\Rightarrow\; s(t) = 1 - \frac{1}{(2\pi)^{1/2}\sigma}\int_0^t \frac{1}{u}\exp\left[-\frac{1}{2}\left(\frac{\log au}{\sigma}\right)^2\right]du$$

Por lo tanto si se toma la función acumulada de una función normal estándar

$$\Phi(x) = \int_{-\infty}^{z} \frac{1}{(2\pi)^{1/2}}\exp[-\frac{1}{2}x^2]dx$$

la función 2.36 se puede escribir como

$$s(t) = 1 - \Phi(\frac{\log au}{\sigma}) \tag{2.37}$$

y de 2.6 se tiene la función de riesgo

$$h(t) = \frac{\dfrac{1}{(2\pi)^{1/2}\sigma t}\exp\left[-\dfrac{1}{2}\left(\dfrac{\log t - \mu}{\sigma}\right)^2\right]}{1 - \Phi\left(\dfrac{\log au}{\sigma}\right)}. \tag{2.38}$$

Observación 7. *La distribución log-normal es muy conveniente por su uso con datos no censurados, ya que la transformación* log *transforma los datos de tal forma que éstos se puedan manejar con un modelo de regresión lineal estándar.*

Es muy común que a diferencia de los modelos exponenciales y Weibull por lo general no tengan formas cerradas para las probabilidades. A diferencia de las dos primeras que si admiten formas cerradas y expresiones sencillas de manejar para la función de riesgo y supervivencia.

Capítulo 3

Tablas de Vida

Las tablas de vida y los métodos que se utilizan para estimarlas han tenido su mayor afluencia en las areas de la actuaría y la investigación médica. Estas reflejan o expresan los datos de supervivencia de tal forma que se permita determinar las probabilidades de riesgo y supervivencia.

3.1. Estructura de una tabla de vida

Se establecen a seguir tres pasos para el análisis de una tabla de vida.

i. Dividir el tiempo de supervivencia en intervalos de longitud bien definida.

ii. Estimar la probabilidad condicional.

iii. Estimar s en los puntos finales.

Supongase que se tiene una muestra de n elementos, los cuales pueden presentar o no el evento esperado en un cierto tiempo t y la duración del estudio, calendarizado, tiene una duración de $(0, \tau)$ donde τ puede ser un tiempo indeterminado o con límite.

Ahora el intervalo $[0, \tau)$ es particionado en secuencias fijas de intervalos $I_1 = (0, t_1], I_2 = (t_1, t_2], \ldots, I_k = (t_{k-1}, \tau]$. Estos intervalos no son necesariamente de longitud igual.

Para la construcción de la tabla de vida sean:
n_i es el número de elementos en riesgo en el comienzo de I_i (es decir, vivos y no censurados)
d_i es el número de eventos que se presentan durante I_i
l_i es el número de pérdidas durante I_i (es decir, elementos que presenta el evento por razones ajenas al estudio)
w_i es el número de salidas del estudio durante I_i (es decir, elementos del estudio que tienen tiempos de supervivencia censurados)

P_i es la probabilidad P(sobrevivieron todo I_i| vivían al comienzo de I_i)
$Q_i = 1 - P_i$ es la P(un elemento presente el evento en I_i| sobrevive más del intervalo I_{i-1})

La cantidad de elementos que se conoce que no han presentado el evento al comienzo de I_i es n_i y entonces para el i-ésimo intervalo, I_1, es $n_1 = n$ donde n es el número total de elementos por lo que para I_i es $n_i = n_{i-1} - d_{i-1} - w_{i-1} \quad , \ i = 2, \ldots, n$ donde d_{i-1} y w_{i-1} son las pérdidas y las salidas respectivamente durante el intervalo I_{i-1}.

Ahora

$$P_i = P(\text{sobrevivió todo } I_i| \text{ vivió al comienzo de } I_i) \tag{3.1}$$

$$\Rightarrow \ P_i = P(T > t_i | T > t_{i-1}) \tag{3.2}$$

$$\Rightarrow \ P_i = \frac{s(t_i)}{s(t_{i-1})} \tag{3.3}$$

dado que $t = 0$ y $s(t_0) = s(0) = 1$ se tiene que

$$P_1 = \frac{s(t_1)}{s(t_0)} = \frac{s(t_1)}{s(0)} = s(t_1)$$

.

Si se evalúa la función de supervivencia en lo punto final del intervalo I_i entonces se obtiene como resultado que $s(t_i)$ queda especificado como:

$$s(t_i) = \left(\frac{s(t_1)}{s(t_0)}\right)\left(\frac{s(t_2)}{s(t_1)}\right)\left(\frac{s(t_3)}{s(t_2)}\right)\cdots\left(\frac{s(t_{i-1})}{s(t_{i-2})}\right)\left(\frac{s(t_i)}{s(t_{i-1})}\right) \tag{3.4}$$

Nótese que en el lado derecho de (3.4) se encuentra la probabilidad condicional definida por (3.3) por lo que la probabilidad de supervivencia se puede descomponer como el producto de probabilidades condicionales:

$$s(t_i) = P_1 P_2 P_3 \cdots P_i; \qquad i = 1, 2, \ldots, n \tag{3.5}$$

Se necesita estimar P_i para que s lo sea en t_i. El estimador más usual para P_i es el "estimador binomial clásico"para una porción:

$$est.\ P_i = 1 - \frac{\#\text{ de elementos que presentaron el evento en } I_i}{\#\text{ de elementos con posibilidad para presentar el evento en } I_i} \tag{3.6}$$

Es decir, un estimador para P_i podría ser $1 - \dfrac{d_i}{n_i}$, si no existiesen elementos censurados.

Pero el problema que se tiene con el denominador de (3.6) es que dicho número se ve afectado o depende de los elementos censurados. Por lo que el tamaño de la *muestra efectiva* es:

$$n_i^{'} = n_i - \frac{1}{2}(l_i + w_i)$$

Si toda la censura ocurriese inmediatamente en el comienzo del intervalo I_i, entonces el número de elementos que pueden presentar el evento estaría dado por $n_i - w_i$; pero si toda la censura se presentase al final de I_i, entonces el número de elementos que pueden presentar el evento estaría dado por n_i. Entonces para calcular el valor del tamaño de la muestra que se busca se usa $\dfrac{(n_i - w_i) + n_i}{2}$.

Y se define el número efectivo de elementos en riesgo, $n_i^{'}$, por

$$n_i^{'} = n_i - \frac{1}{2} w_i$$

Esta consideración determina que la censura ocurre uniformemente en el intervalo, en consecuencia, el estimador de P_i, $\hat{P}_i$ referido como actuarial, está dado por

$$\hat{P}_i = 1 - \frac{d_i}{n_i^{'}} \;\Rightarrow\; \hat{Q}_i = \frac{d_i}{n_i^{'}} \qquad i = 1, 2, \ldots, n$$

Observación 8. *Si para un intervalo I_r, $n_r^{'} = 0$ entonces $\hat{P}_i$ el estimador actuarial toma el valor de cero, $\hat{P}_i = 0$.*

Por lo tanto el llamado estimador actuarial de supervivencia $\hat{s}(t_i)$ está dado por

$$\hat{s}(t_i) = \hat{P}_1 \hat{P}_2 \hat{P}_3 \cdots \hat{P}_i = \prod_{r=1}^{i} \hat{P}_r; \qquad i = 1, \ldots, n,\ t_i = t_0, t_1, \ldots, t_n \tag{3.7}$$

Por lo tanto la estructura de una tabla de vida se determina como se señala acontinuación

I_i Intervalo	d_i Muertos	l_i Perdidas (desconocidos)	w_i Salidas (censurados)	n_i En riesgo	$n_i^{'}$ (# efectivo en riesgo)	$\hat{Q}_i$	$\hat{P}_i$	$\hat{s}(t_i)$

Tabla 3.1: Estructura de una tabla de vida

3.2. Fórmula de Greenwood

No sería correcto que al momento de estimar la tabla de vida se pensara que ésta no contenga errores de estimación, es por eso que surge la necesidad de que se calcule el error estándar. Y el modo más recomendable es usar la fórmula de Greenwood para establecer aproximadamente la varianza de $s(t_i)$.

Teorema 1. *Sea $t_i, \quad i = 1, \cdots, n$ variables de tiempos observados y P_i como se definió en la sección anterior $E(\hat{s(t_i)}) \cong P_1 P_2 P_3 \cdots P_i = s(t_i)$; $i = 1, 2, \ldots, n$*

Demostración.
Sea $r < i$ entonces sobre la condicional $n'_r > 0$, por lo que se tiene un conteo binomial. Nótese que $d_i|n'_i \sim Bin(n'_i, 1 - P_i)$.
$\Rightarrow$ $E(1 - \hat{P}_i)$ usando esperanza condicional:

$$\begin{aligned} E(1-\hat{P}_i) &= E(\hat{Q}_i) \\ &= E\Big(\frac{d_i}{n'_i}\Big) \\ &= E\Big[E\Big(\frac{d_i}{n'_i}|n'_i\Big)\Big] \\ &= E\Big[\frac{1}{n'_i}E\Big(d_i|n'_i\Big)\Big] \\ &= E\Big[\frac{1}{n'_i}(n'_i(1-P_i))\Big] \\ &= E(1-P_i) \\ &= 1-P_i \end{aligned}$$

$$\therefore\ E(1-\hat{P}_i) = 1-P_i$$

$$\Rightarrow\ E(\hat{P}_i) \cong P_i$$

de igual manera en función de que $n'_i > 0$, se tiene que para $r < i$

$$E(\hat{P}_r\hat{P}_i) \cong P_rP_i$$

$$\begin{aligned} \Rightarrow\ E(\hat{s}(t_i))) &= E(\hat{P}_1\hat{P}_2\cdots\hat{P}_i) \\ &\cong P_1P_2\cdots P_i \end{aligned}$$

esta aproximación es precisa debido a que $P(n'_i = 0)$ es pequeña.

$$\therefore\ E(\hat{s}) \cong P_1P_2P_3\cdots P_i = s(t_i)\ i = 1, 2, \ldots, n$$

Teorema 2. *Fórmula de Greenwood*
El error estándar de la tabla de vida estimada en un punto fijo t_i está dado por:

$$var(\hat{s}(t_i)) \cong [s(t_i)]^2 \sum_{r=1}^{i} \frac{Q_r}{P_r n'_r}\quad i = 1, 2, \ldots, n$$

Demostración.
Sea $j < i$ entonces sobre la condicional $n'_i > 0$, desde que los datos siguen una distribución binomial

$$d_i|n'_i \sim Bin(n'_i, 1 - P_i)$$

Por el teorema anterior se tiene que la $var(\hat{P}_i) = \dfrac{P_i(1-P_i)}{n'_i}$ pues

$$E(\hat{P}_i^2) \cong \frac{P_i(1-P_i)}{n'_i} + P_i^2 = P_i^2\Big(1 + \frac{Q_i}{P_i n'_i}\Big)$$

$$\begin{aligned}
\Rightarrow\ var(\hat{s}(t_i)) &= E[\hat{s}(t_i)^2] - [E[\hat{s}(t_i)]]^2 \\
&\approx \prod_{r=1}^{i} P_r^2\Big(1 + \frac{Q_r}{P_r n'_r}\Big) - (s(t_i))^2 \\
&= (s(t_i))^2\Big[\prod_{r=1}^{i}\Big(1 + \frac{Q_r}{P_r n'_r}\Big) - 1\Big] \\
&\cong (s(t_i))^2\Big[1 + \sum_{r=1}^{i}\frac{Q_r}{P_r n'_r} - 1\Big] \\
&= (s(t_i))^2\Big[\sum_{r=1}^{i}\frac{Q_r}{P_r n'_r}\Big]
\end{aligned}$$

donde, en la penúltima línea, el término de orden $\dfrac{1}{(n'_r)^2}$ crece rápidamente por lo cual se tiene que:

$$\prod_{r=1}^{i}(1 + a_r) \approx 1 + \sum_{r=1}^{i} a_r \text{ donde } a_r = \frac{Q_r}{P_r n'_r}$$

$$\therefore\ var(\hat{s}(t_i)) \cong [s(t_i)]^2 \sum_{r=1}^{i}\frac{Q_r}{P_r n'_r} \quad i = 1, 2, \ldots, n$$

Dado que $P_i = 1 - \dfrac{d_i}{n'_i}$ y $Q_i = \dfrac{d_i}{n'_i}$ la fórmula de Greenwood queda determinada como:

$$est.\ var[\hat{s}(t_i)] = [\hat{s}(t_i)]^2 \sum_{r=1}^{i}\frac{d_r}{n'_r(n'_r - d_r)} \tag{3.8}$$

Del proceso de estimación de $\hat{s}(t_i)$ y de su varianza se tiene que el error estándar está dado por:

$$\begin{aligned}
E.E.\Big[\hat{s}(t_i)\Big] &= \hat{s}(t_i)\sqrt{\sum_{r=1}^{i}\frac{Q_r}{P_r n'_r}} \ ; \quad i = 1, 2, \ldots, n \\
&= \hat{s}(t_i)\sqrt{\sum_{r=1}^{i}\frac{d_r}{n'_r(n'_r - d_r)}} \ ; \quad i = 1, 2, \ldots, n
\end{aligned}$$

Ejemplo 5. *Tabla de vida.*

Los datos son de 913 hombres y mujeres, pacientes con melanoma pernicioso tratados en la clínica del tumor, M.D. Anderson entre 1944 y 1954c, el estudio fue empleado por Gross y Clark (1975). Tabla 3.2.

I_i **Intervalo (años)**	d_i **Muertes**	w_i **Pérdidas**	n_i **En riesgo**	n_i' **# efectivo en riesgo**	$\widehat{Q}_i$	$\hat{P}_i$	$\hat{s}(t_i)$
[0,1)	312	96	913	865	0.361	0.639	0.639
[1,2)	96	74	505	468	0.205	0.795	0.508
[2,3)	45	62	335	304	0.148	0.852	0.433
[3,4)	29	30	228	213	0.136	0.864	0.374
[4,5)	7	40	169	149	0.047	0.953	0.356
[5,6)	9	37	122	103.5	0.087	0.913	0.325
[6,7)	3	17	76	67.5	0.044	0.956	0.311
[7,8)	1	12	56	50	0.020	0.980	0.305
[8,9)	3	8	43	39	0.077	0.923	0.281
[9,∞)	32	-	32	32	1.000	0.000	0.000

Tabla 3.2: Pacientes con melanoma pernicioso

se realiza el calculo en el intervalo $[5, 6)$ solo para ejemplificar los datos en la tabla (3.2)

$$\hat{Q} = \frac{9}{103{,}5} = 0.087\ ; \quad \hat{P} = 1 - \hat{Q} = 0.913$$

$$\Rightarrow\ est.\ var[\hat{s}(5)] = 0.00039\ \Rightarrow\ \ E.E.[\hat{s}(5)] = 0.01996$$

En la **ausencia de datos censurados** el estimador de $s(t)$ como ya se ha visto está dado por

$$\hat{s}(t) = s_n(t) = \frac{\#\text{ de elementos después de } t}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{(t,\infty)}(T_i)$$

donde T_i son n variables i.i.d. con distribución T y $s(t) = P(T > t)$. Para un valor fijo t, $s_n(t)$ es un promedio de n como antes

$$\Rightarrow \; ns_n(t) = \sum_{i=1}^{n} I_{(t,\infty)}(T_i)$$

Nótese que cada $I_{(t,\infty)}(T_i)$ tiene una distribución Bernuolli, es decir,

$$P(I_{(t,\infty)}(T_i) = 1) = P(T_i > t) = s(t)$$

$$P(I_{(t,\infty)}(T_i) = 0) = P(T_i < t) = 1 - s(t)$$

por lo que $E\Big[I_{(t,\infty)}(T_i)\Big] = s(t)$ y $var\Big[I_{(t,\infty)}(T_i)\Big] = s(t)(1 - s(t))$

Por lo que se tiene que $s_n(t) \sim Bin(n, s(t))$, calculando la esperanza y la varianza en un punto fijo t se tiene que:

$$E[s_n(t)] = \frac{1}{n}(ns(t)) = s(t)$$

$$var[s_n(t)] = \frac{1}{n^2}[ns(t)(1 - s(t))] = \frac{s(t)(1 - s(t))}{n}$$

Por lo tanto, cuando no hay presencia de datos censurados el error estándar es representado por:

$$E.E.[\hat{s}(t)] = \sqrt{\frac{\hat{s}(t)(1 - \hat{s}(t))}{n}}$$

y se sigue una aproximación normal se puede obtener un intervalo de confianza al $(1 - \alpha)100\,\%$, para una distribucion normal.

$\hat{s}(t) \pm Z_{1-\frac{\alpha}{2}} E.E.[\hat{s}(t)]$ donde $Z_{1-\frac{\alpha}{2}}$ es el percentil de una distribución normal estándar.

Este intervalo es simétrico y próximo al punto estimado, sin embargo, asimétrico excepto para $\hat{s}(t)$ alrededor de 0.5 por lo que la simetría no podría ser del todo exacta. Aunque en teoría, en el orden para representar la asimetría de la distribución de la muestra, los límites deberían ser asimétricos en el contorno de $\hat{s}(t)$ y en un rango admisible. Este hecho podría ser seguido como un recurso para la estadística:

$$Z = \frac{\hat{s}(t) - s(t)}{\sqrt{\frac{s(t)[1 - s(t)]}{n}}}$$

Por lo tanto los límites de confianza a $(1 - \alpha)100\,\%$ para $s(t)$ queda determinado por $[s_{inf}(t), s_{sup}(t)]$, donde s_{inf} y s_{sup} son el límite inferior y superior respectivamente tal que

$$s_{inf}(t) = \frac{\left(2n\hat{s}(t) + Z^2_{1-\frac{\alpha}{2}}\right) - Z_{1-\frac{\alpha}{2}}\left[4n\hat{s}(t)(1 - \hat{s}(t)) + Z^2_{1-\frac{\alpha}{2}}\right]}{2(n + Z^2_{1-\frac{\alpha}{2}})},$$

$$s_{sup}(t) = \frac{\left(2n\hat{s}(t) + Z^2_{1-\frac{\alpha}{2}}\right) + Z_{1-\frac{\alpha}{2}}\left[4n\hat{s}(t)(1-\hat{s}(t)) + Z^2_{1-\frac{\alpha}{2}}\right]}{2(n + Z^2_{1-\frac{\alpha}{2}})}$$

3.3. Estimador Kaplan-Meier

3.3.1. Método de máxima verosimilitud

Antes de encontrar el estimador Kaplan-Meier se describirá a continuación el método de máxima verosimilitud el cual nos permite encontrar un la estimación de algún parámetro más probable nuestra muestra.

Sean $Y_1, Y_2, \cdots, Y_n$ variables aleatorias con f.d.p. $f(\underline{Y};\underline{\theta})$ la cual depende de $\underline{\theta}$, donde $\underline{Y} = (Y_1, Y_2, \cdots, Y_n)$ y $\underline{\theta} = (\theta_1, \cdots, \theta_r)$.

La función de verosimilitud $L(\underline{\theta};\underline{Y})$ donde la notación hace énfasis de los parámetros de $\underline{\theta}$ con $\underline{Y}$ fijo (donde $\underline{Y}$ representa las observaciones).

Sea Ω que denota el conjunto de todos los posibles valores de $\underline{\theta}$. El estimador de máxima verosimilitud (E.M.V.) de $\underline{\theta}$ es el valor $\hat{\underline{\theta}}$ el cual maximiza la función de verosimilitud, esto es

$$L(\hat{\underline{\theta}};\underline{Y}) \geq L(\underline{\theta};\underline{Y}) \quad \forall \underline{\theta} \in \Omega$$

Además, una característica muy importante de $\hat{\underline{\theta}}$ es que también maximiza la función de verosimilitud logarítmica, $\lambda(\underline{\theta};\underline{Y}) = \log L(\underline{\theta};\underline{Y})$, en virtud de que la función logaritmo es monótona y creciente.

$$\Rightarrow \ \lambda(\hat{\underline{\theta}};\underline{Y}) \geq \lambda(\underline{\theta};\underline{Y}) \quad \forall \underline{\theta} \in \Omega$$

Esta característica se resalta por el hecho de que resulta más sencillo trabajar con el logaritmo de la función de verosimilitud.

$\hat{\underline{\theta}}$ se obtiene por diferenciación y resolviendo las ecuaciones correspondientes para cada θ_j, por medio de la función logarítmica:

$$\frac{d\lambda(\underline{\theta};\underline{Y})}{d\theta_j} = 0 \text{ para } j = 1, 2, \cdots, r$$

Luego se muestra que la matriz de segundas derivadas de $\lambda(\underline{\theta};\underline{Y})$ que está dada por

$$\frac{d^2\lambda(\underline{\theta};\underline{Y})}{d\theta_j d\theta_k} = \begin{pmatrix} \frac{d^2\lambda(\underline{\theta};\underline{Y})}{d\theta_1 d\theta_1} & \frac{d^2\lambda(\underline{\theta};\underline{Y})}{d\theta_1 d\theta_2} & \cdots & \frac{d^2\lambda(\underline{\theta};\underline{Y})}{d\theta_1 d\theta_r} \\ \frac{d^2\lambda(\underline{\theta};\underline{Y})}{d\theta_2 d\theta_1} & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \frac{d^2\lambda(\underline{\theta};\underline{Y})}{d\theta_r d\theta_1} & \cdots & \cdots & \frac{d^2\lambda(\underline{\theta};\underline{Y})}{d\theta_r d\theta_r} \end{pmatrix}$$

Evaluada en $\underline{\theta} = \hat{\underline{\theta}}$ está debe estar definida negativa. Y entonces si un parámetro θ verifica que $\dfrac{d^2\lambda(\underline{\theta};\underline{Y})}{d\theta^2}$ evaluado en $\theta = \hat{\theta}$ es negativo el estimador $\hat{\underline{\theta}}$ se dice que es de máxima verosimilitud.

Observación 9. *Si $g(\underline{\theta})$ es una función arbitraria del parámetro $\underline{\theta}$, entonces $g(\hat{\underline{\theta}})$ es el estimador de máxima verosimilitud de $g(\underline{\theta})$. Esta propiedad se define como "de invarianza"del E.M.V.*

3.3.2. El estimador producto-límite

Kaplan y Meier (1958) consideraron estudiar el estimador producto límite de una distribución de supervivencia sin la necesidad de asumir cualquier forma paramétrica en particular, y es de aquí de donde se considera no paramétrico.

El estimador producto-límite, se desarrolló para estimar la función de supervivencia $s(t)$, usando por convención datos con presencia de censura por la derecha y asignando como guía las hipótesis del estimador actuarial, aunque a diferencia de este último en él la longitud del intervalo no es fija y los datos pueden ser censurados o no.

Sean $T_1, T_2, \ldots, T_n$ los tiempos de supervivencia de los n elementos de la muestra a estudiar. Ahora con cada T_i se asocia una variable C_i, conocida como variable de censura, la cual es registrada cuando el tiempo en que se presenta el evento es desconocido. Se considera a

$$Z_i = min(T_i, C_i) \text{ y } \delta_i = \begin{cases} 1 \text{ si } T_i \leq C_i \text{ (No censurados)} \\ 0 \text{ si } T_i > C_i \text{ (Censurados)} \end{cases}$$

Por lo que se tienen las parejas (Z_i, δ_i).

Esta consideración se da por el hecho de que los datos ocurren naturalmente como pares ordenados, dado que se sabe cual de ellos son censurados y cuales no.

Se asume inicialmente que las observaciones son independientes.

Sean $Z_{(1)} < Z_{(2)} < \cdots < Z_{(n)}$ las estadísticas de orden de $Z_1, Z_2, \cdots, Z_n$ y los datos observados son todos diferentes.

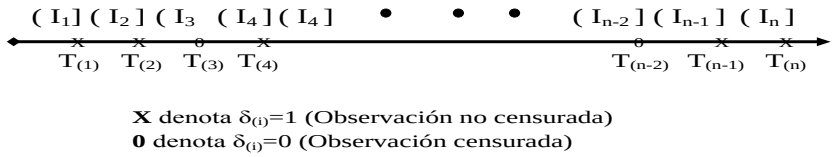
Además se debe de garantizar cuando los elementos observados son ordenados de menor a mayor. Por lo que se toma a $\delta_{(i)}$ como el valor del indicador asociado con $Z_{(i)}$, el cual ha sido nuevamente definido, es decir, $\delta_{(i)} = \delta_j$ si $Z_{(i)} = Z_j$. Por lo que se asume que la variable aleatoria siendo medida es independiente de los tiempos en que se genera la censura.
Sean:

$R(t)$ el conjunto en riesgo en el tiempo t, es decir, es el grupo de los elementos que se encuentran vivos en t^- (exactamente antes de t).
n_i es el número de elementos en riesgo $R(Z_{(i)})$ representa al número de elementos que no

han presentado el evento en $Z_{(i)}^{-}$.
d_i =número de elementos que presentaron el evento esperado en $Z_{(i)}$.

Nótese que para los datos observados dado que no están ligados $d_i = 1$ o $d_i = 0$ que dependen de si $\delta_{(i)} = 1$ o $\delta_{(i)} = 0$.
P_i es la probabilidad P[número de elementos que no presentaron el evento en todo I_i| estaban presentes al comienzo de I_i], es decir, $P_i = P[T > Z_{(i)}|T > Z_{(i-1)}]$, por lo tanto $Q_i = 1 - P_i$

Donde I_i es un subintervalo del intervalo del tiempo ya calendarizado de interés, $(0, Z_{(i)}]$, el cual es dividido en n subintervalos I_j con punto final $Z_{(i)}$.



Gráfica 3.1: Intervalo calendario, seccionado.

n_i y el $R(t)$ son usados para estimar P_i, por lo que el estimador propuesto es:

$$\begin{aligned}
\hat{P}_i &= 1 - \frac{\#\text{ de elementos que presentan el evento en } (0, Z_{(i)}]}{\#\text{ de elementos con potencial de presentar el evento en } (0, Z_{(i)}]} \\
&= 1 - \frac{d_i}{n_i} \\
&= \begin{cases} 1 - \dfrac{1}{n_i}, & \text{si } \delta_{(i)} = 1 \text{ (No censurados)} \\ 1, & \text{si } \delta_{(i)} = 0 \text{ (Censurados)} \end{cases}
\end{aligned}$$

Se retoma el hecho de que $s(t_i) = P_1 P_2 \cdots P_i$, $i = 1, 2, \cdots, n$ y además nótese que como i se incrementa, el conjunto de riesgo disminuye uno en un tiempo, por lo que se tiene que

$n_i = n - (i-1) = n - i + 1$.

$$\begin{aligned}
\Rightarrow\ \hat{s}(t) &= \prod_{i;Z_{(i)}\leq t} \hat{P}_i \\
&= \prod_{i;Z_{(i)}\leq t,\delta_{(i)}} \Big(1 - \frac{1}{n_i}\Big) \\
&= \prod_{i;Z_{(i)}\leq t,} \Big(1 - \frac{1}{n_i}\Big)^{\delta_{(i)}} \\
&= \prod_{i;Z_{(i)}\leq t,} \Big(1 - \frac{1}{n-i+1}\Big)^{\delta_{(i)}} \\
&= \prod_{i;Z_{(i)}\leq t,} \Big(\frac{n-i}{n-i+1}\Big)^{\delta_{(i)}}
\end{aligned}$$

Por lo tanto si t es fijado y no hay posibilidad de que las observaciones censuradas por la derecha estén ligadas

$$(Z_{(1)}, \delta_{(1)}), (Z_{(2)}, \delta_{(2)}), \ldots, (Z_{(n)}, \delta_{(n)})$$

entonces el estimador producto-límite de $s(t)$ es definido por

$$\hat{s}(t) = \prod_{i;Z_{(i)}\leq t,} \Big(\frac{n-i}{n-i+1}\Big)^{\delta_{(i)}}.$$

Ahora considérese el hecho de que la censura es aleatoria, la que por convención se trabaja por la derecha, y sin la necesidad de que se asuma que existen observaciones dependientes, ya sean éstas censuradas o no. Y nótese que las ques están relacionadas a dos valores (t_i, δ_i), obtenidos de dos procesos por separado, el del evento en un tiempo t y en el que se censura C. Ya establecida la hipótesis la censura con que se trabaja, la función de verosimilitud de la muestra de tamaño n es

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} s(t_i)^{1-\delta_i} \tag{3.9}$$

donde $f(t)$ está determinada por los elementos que presenta el evento esperado (falla) para la probabilidad alrededor del tiempo en el que se espera el evento (falla) y las observaciones censuradas determinan $s(t)$ al rededor del tiempo de censura.

Considerando la muestra de n elementos a observar (t_i, δ_i) $i = 1, 2, \ldots, n$, en los cuales los eventos son observados en r distintos tiempos, ya que se pueden repetir observaciones en un mismo tiempo, considérese las estadísticas de orden de t_i, $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$, y que la distribución de supervivencia es una función de paso en los tiempos de los eventos.

El límite por la izquierda $F(t^-) = \lim_{\Delta \to 0} F(t - \Delta) = \lim_{u \to t} F(u)$ determina el valor de la función de distribución acumulada antes de presentarse un salto o la discontinuidad en el tiempo t, entonces la probabilidad de que un evento se presente en un tiempo $t_{(i)}$ es

$$\begin{aligned} f(t_{(i)}) &= F(t_{(i)}) - F(t_{(i)}^-) \\ &= 1 - s(t_{(i)}) - [1 - s(t_{(i)}^-)] \\ &= s(t_{(i)}^-) - s(t_{(i)}) \end{aligned}$$

donde $s(t_{(i)})^- = P(T \geq t_{(i)})$ es la probabilidad de estar en riesgo en $t_{(i)}$ y $s(t_{(i)}) = P(T > t_{(i)})$ es la probabilidad de no presentar el evento más alla de t_i. Desde que $f(t_{(i)})$ es la probabilidad de supervivencia en t_i y que ocurra el evento en t_i, se tiene que

$$f(t_{(i)}) = P[\text{el evento ocurra en } t_{(i)}]P[T \geq t_{(i)}]$$

$$\Rightarrow \; f(t_{(i)}) = Q_i s(t_{(i)})^-$$

donde $Q_i = \lim_{\Delta t \to 0} P[t_{(i)}^- < T \leq t_{(i)}^- + \Delta t | T \geq t_{(i)}]$

$$\begin{aligned} 1 - Q_i &= P_i \\ &= P[T > t_{(i)} | T > t_{(i)}^-] \\ &= P[T > t_{(i)} | T \geq t_{(i)}^-] \end{aligned}$$

dado que $s(t)$ es una función decreciente, se tiene que

$$f(t_{(i)}) = Q_i s(t_{(i)}^-) = Q_i s(t_{(i-1)})$$

sean:
n_i es el número de elementos en riesgo en $t_{(i)}$
d_i es el número de eventos que se presentan en $t_{(i)} = \{t_i = t_{(j)} \delta_i = 1\}$
w_i es el número de eventos censurados después del i-ésimo pero antes del $(i+1)$-ésimo evento que es igual número $\{t_i, \delta_i = 0\} \in [t_{(i)}, t_{(i+1)}]$.

Como se conoce al número de elementos que no han presentado el evento en t_i, se tiene que $n_{i+1} = n_i - d_i - w_i$, entonces las observaciones que son censuradas por la derecha en el tiempo del evento $t_{(i)}$ son consideras por estar en riesgo en aquel tiempo y son quitadas del conjunto de riesgo posterior inmediatamente. Por lo que la función de verosimilitud puede ser expresada como

$$L(Q_1, \ldots, Q_n) = \prod_{i=1}^{r} Q_i^{d_i} s(t_{(i-1)})^{d_i} s(t_{(i)})^{w_i}; \qquad r = 1, \ldots, n \tag{3.10}$$

donde $t_{(0)} = 0$ y $s(t_{(0)}) = 1$.

dentro del proceso el tiempo en el cual se presenta el evento más allá de $t_{(i)}$ da por hecho que sobrevivió más allá del tiempo $t_{(i-1)}$, $t_{(i-2)}$, etc.

$$\begin{aligned} \Rightarrow\ s(t_{(i)}) &= P[T > t_{(i)}] \\ &= P[T > t_{(i)}|T > t_{(i-1)}]P[T > t_{(i-1)}|T > t_{(i-2)}] \\ &\quad \cdots P[T > t_{(2)}|T > t_{(1)}]P[T > t_{(1)}|T > t_{(0)}] \end{aligned}$$

Nótese que

$$\begin{aligned} P[T > t_{(i)}|T > t_{(i-1)}] &= P[T > t_{(i)}|T > t^{-}_{(i)}] \\ &= P[T > t_{(i)}|T \geq t_{(i)}] \\ &= P_i \end{aligned}$$

$$\begin{aligned} \Rightarrow\ s(t_{(i)}) &= P_i s(t_{(i-1)}) \\ &= P_i P_{i-1} P_{i-2} \cdots P_2 P_1 \end{aligned}$$

por lo que sólo basta con que se maximice

$$\begin{aligned} L(Q_1, \ldots, Q_r) &= \prod_{i=1}^{r} Q_i^{d_i} (P_1 P_2 P_3 \cdots P_{i-1})^{d_i} (P_1 P_2 P_3 \cdots P_i)^{w_i} \\ &= \prod_{i=1}^{r} Q_i^{d_i} P_i^{n_i - d_i} \end{aligned}$$

con la función de verosimilitud logarítmica

$$\lambda(Q_1, \ldots, Q_r) = \sum_{i=1}^{r} [d_i \log(Q_i) + (n_i - d_i) \log(P_i)]) \tag{3.11}$$

usando el método de máxima verosimilitud para hallar el estimador se tiene que este ultimo se obtiene de las ecuaciones

$$\frac{d\lambda(Q_1, \ldots, Q_r)}{dQ_i} = \frac{d_i}{Q_i} - \frac{n_i - d_i}{P_i} = 0 \tag{3.12}$$

$$\Rightarrow\ \frac{d_i}{Q_i} - \frac{n_i - d_i}{1 - Q_i} = 0$$

$$\therefore\ \hat{Q}_i = \frac{d_i}{Q_i}\ \Rightarrow\ \hat{P}_i = \frac{n_i - d_i}{n_i}$$

Por lo tanto el estimador de máxima verosimilitud mostrado por Kaplan y Meier, en cual si cualquier elemento es censurado en el tiempo $t_{(i)}$ entonces es considera para poder sobrevir por más tiempo que las fallas en $t_{(i)}$, es decir, se considera en riesgo y se define como:

$$\hat{s}(t) = \prod_{i=1;\ t_i \leq t}^{r} \left(\frac{n_i - d_i}{n_i}\right) \tag{3.13}$$

3.4. Estimación de la función de riesgo

Supongase que los tiempos de supervivencia observados son agrupados en k intervalos de la misma forma en que se construyeron en la tabla de vida. Una forma de estimar la función de riesgo es que se hace uso de los datos; se toma el rango del número de fallas en un tiempo dado para la cantidad de elementos en riesgo en dicho tiempo. Se asume que la función de riesgo es constante entre los sucesivos tiempos de falla, el riesgo en un tiempo requerido se puede encontrar si se divide por el intervalo de tiempo.

Si hay d_i fallas en el tiempo i-ésimo, t_i $i = 1, 2, \ldots, k$, y n_i observaciones (o elementos) en riesgo de presentar la falla en el tiempo t_i, la función de riesgo en el intervalo $[t_i, t_{i+1})$ se puede estimar por:

$$\hat{h}(t) = \frac{d_i}{n_i \tau_i} \quad \text{donde}\ \ t_i \leq t < t_{i+1}\ \ \text{y}\ \ \tau_i = t_{i+1} - t_i \tag{3.14}$$

El error se puede encontrar desde la varianza de d_i con una distribución binomial con parámetros n_i y P_i asumida; recuérdese que P_i es la probabilidad de falla en el intervalo de longitud τ_i y por el teorema (1) se tiene

$$var(d_i) = n_i \hat{P}_i (1 - \hat{P}_i)\ \ \text{y}\ \ \hat{P}_i = 1 - \frac{d_i}{n_i} = \frac{n_i - d_i}{n_i}$$

$$
\begin{aligned}
var[\hat{h}(t)] &= var\Big(\frac{d_i}{n_i\tau_i}\Big) \\
&= \Big(\frac{1}{n_i^2\tau_i^2}\Big)var(d_i) \\
&= \Big(\frac{1}{n_i^2\tau_i^2}\Big)n_i\hat{P}_i(1-\hat{P}_i) \\
&= \Big(\frac{1}{n_i^2\tau_i^2}\Big)n_i\Big(1-\frac{d_i}{n_i}\Big)\Big(\frac{d_i}{n_i}\Big) \\
&= \Big(\frac{1}{n_i^2\tau_i^2}\Big)\Big(d_i-\frac{d_i^2}{n_i}\Big) \\
&= \Big(\frac{d_i^2}{n_i^2\tau_i^2}\Big)\Big(\frac{1}{d_i}-\frac{1}{n_i}\Big) \\
&= [\hat{h}(t)]^2\Big(\frac{1}{d_i}-\frac{1}{n_i}\Big)
\end{aligned}
$$

$$
\therefore \quad var[\hat{h}(t)] = \hat{h}^2(t)\Big(\frac{n_i-d_i}{n_id_i}\Big) \tag{3.15}
$$

$$
\therefore \quad E.E.[\hat{h}(t)] = \hat{h}(t)\left(\sqrt{\frac{n_i-d_i}{n_id_i}}\right) \tag{3.16}
$$

El intervalo de confianza para $h(t)$ con un $(1-\alpha)100\,\%$ de confianza, está dado por

$$
\hat{h}(t) \pm Z_{1-\frac{\alpha}{2}}E.E.[\hat{h}(t)]
$$

donde $Z_{1-\dfrac{\alpha}{2}}$ es el cuantil del $(1-\dfrac{\alpha}{2})100\,\%$ de una distribución normal estándar.

3.4.1. Estimación de la función de riesgo acumulada

La función de riesgo acumulada es importante en la identificación de modelos con datos de supervivencia ya que su derivada es la de riesgo, además de dar información de la supervivencia.

La función de riesgo acumulada en el tiempo t, que se denota como $H(t)$, fue definida en la ecuación (2.8), pero debido a que esto se considera una forma difícil de calcular lo que se hace es trabajar con la función definida por (2.9)

$$
H(t) = -\log[s(t)]
$$

Así si se hace uso del estimador Kaplan-Meier, (3.13), $\hat{H}(t) = -\log[\hat{s}(t)]$ es un estimador natural apropiado en el tiempo t:

$$
\hat{H}(t) = -\log\left[\prod_{i=1;\, t_i\leq t}^{r}\Big(\frac{n_i-d_i}{n_i}\Big)\right] = -\sum_{i=1;\, t_i\leq t}^{r}\log\Big(\frac{n_i-d_i}{n_i}\Big)
$$

para $t_{(i)} \leq t < t_{r+1}$, $r = 1, 2, \ldots, n$, y $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$ son los r tiempos de falla ordenados con $t_{(r+1)} = \infty$.

Un estimador alterno para la función de riesgo acumulada fue presentado por Nelson y Alen el cual está basado en los tiempos de los eventos individuales y es t_i dado por

$$\hat{H}(t) = \sum_{i=1;\, t_i \leq t}^{r} \left(\frac{d_i}{n_i}\right) \tag{3.17}$$

el cual es la suma acumulada de las probabilidades estimadas de falla desde el primero hasta el k-ésimo tiempo de falla.

Ahora se toma el hecho de que $\hat{s}(t) = \prod_{i=1;\, t_i \leq t}^{r} \hat{P}_i \quad r = 1, 2, \ldots, n$ donde $\hat{P}_i = \left(\frac{n_i - d_i}{n_i}\right)$

$$\Rightarrow \log[\hat{s}(t)] = \sum_{i=1;\, t_i \leq t}^{r} \log \hat{P}_i$$

por el hecho de que los r tiempos son distintos y las P_i son independientes

$$\Rightarrow var\Big[\log[\hat{s}(t)]\Big] = \sum_{i=1;\, t_i \leq t}^{r} var[\log \hat{P}_i]$$

como $\hat{P}_i = \left(\frac{n_i - d_i}{n_i}\right) \Rightarrow var[\hat{P}_i] = \frac{\hat{P}_i(1 - \hat{P}_i)}{n_r}$

Se retoma el resultado de la varianza de la función $g(x)$ en términos de la variable aleatoria x, este resultado se conoce como *la aproximación de la serie de Taylor* o *método delta*, el cual está dado por

$$var[g(x)] = \left[\frac{dg(x)}{dx}\right]^2 var[x] \tag{3.18}$$

se tiene que $var[\log \hat{P}_i] = \frac{(1 - \hat{P}_i)}{n_i \hat{P}_i}$

$$\therefore var[\log \hat{P}_i] = \frac{d_i}{n_i(n_i - d_i)}$$

$$\therefore var[\log \hat{s}(t)] = \sum_{i=1}^{r} \frac{d_i}{n_i(n_i - d_i)} \tag{3.19}$$

considerando una vez más (3.18) se tiene que

$$\begin{aligned} var[\hat{H}(t)] &= var\Big[-\log(\hat{s}(t))\Big] \\ &= \left(\frac{1}{\hat{s}(t)}\right)^2 var(\hat{s}(t)) \\ &= \left(\frac{1}{\hat{s}^2(t)}\right)\hat{s}^2(t)\sum_{i=1}^{r}\frac{d_i}{n_i(n_i-d_i)} \end{aligned}$$

$$\therefore\ var[\hat{H}(t)] = \sum_{r=1}^{k}\frac{d_r}{n_r(n_r-d_r)}$$

Es importante señalar que la fórmula de Greenwood permite construir intervalos de confianza simétricos, aunque sus extremos no se encuentran delimitados por el intervalo $(0,1)$. Los límites del intervalo de confianza, para que ya no estén delimitados, se hace uso de la transformación complementaria log(log) de la función de riesgo acumulada con varianza estimada correspondiente, se tiene que:

$$var\Big[\log(\hat{H}(t))\Big] = var\Big[\log(-\log[\hat{s}(t)])\Big]$$

y aplicando (3.18) a lo anterior

$$var\Big[\log(-\log[\hat{s}(t)])\Big] = \left(\frac{1}{(\log[\hat{s}(t)])^2}\right)\sum_{i=1}^{r}\frac{d_i}{n_i(n_i-d_i)} \tag{3.20}$$

$$\Rightarrow\ E.E.[\log(-\log[\hat{s}(t)])] = \left(\frac{1}{(\log[\hat{s}(t)])}\right)\sqrt{\sum_{i=1}^{k}\frac{d_i}{n_i(n_i-d_i)}}$$

El resultado del uso de la transformación log(log) permite construir un intervalo de confianza asimétrico con un $(1-\alpha)100\,\%$ de confianza para $s(t)$ usando la estadistíca

$$Z = \frac{\log(-\log[s(t)]) - \log(-\log[\hat{s}(t)])}{E.E.[\log(-\log[\hat{s}(t)])]}$$

por lo que resulta el intervalo de confianza para $s(t)$ de la forma:

$$\hat{s}(t)^{\exp\left[\pm Z_{1-\frac{\alpha}{2}}E.E.[\log(-\log[\hat{s}(t)])]\right]}$$

donde $Z_{1-\frac{\alpha}{2}}$ es el cuantil $(1-\alpha)100\,\%$ con una distribución normal estándar.

Este método para encontrar intervalos asimétricos se puede utilizar para encontrar los límites del intervalo de confianza de la función de riesgo acumulada.

3.5. Prueba para dos grupos con datos de supervivencia

En ocasiones se necesita comparar las distribuciones del tiempo de falla entre dos grupos independientes formados por datos de supervivencia. Se considera la comparación de estos en un tiempo específico haciendo uso del estimador Kaplan-Meier.

La estimación de la fórmula de Greenwood facilita un intervalo de confianza para muestras grandes de la diferencia entre las funciones de supervivencia estimadas de los dos grupos correspondientemente en un tiempo $t > 0$.

Considerese a $\hat{s}_1(t)$ y $\hat{s}_2(t)$ las funciones de supervivencia estimadas con el estimador Kaplan-Meier para los grupos A y B respectivamente (esto con el fin de saber si la supervivencia de un grupo puedes saberse por mediante el otro), entonces

$$\hat{s}_1(t) - \hat{s}_2(t) \;\Rightarrow\; var[\hat{s}_1(t) - \hat{s}_2(t)] = var[\hat{s}_1(t)] + var[\hat{s}_2(t)]$$

Mientras que los límites del intervalo de confianza asimétrico sobre el rango de las probabilidades de supervivencia en el tiempo t se obtienen usando la transformación logarítmica por lo que se tiene que

$$\log(\hat{s}_1(t)) - \log(\hat{s}_2(t)) = \log \frac{\hat{s}_1(t)}{\hat{s}_2(t)}$$

$$\Rightarrow \quad var[\log(\hat{s}_1(t)) - \log(\hat{s}_2(t))] = var[\log(\hat{s}_1(t))] + var[\log(\hat{s}_2(t))]$$

donde $var[\log(\hat{s}_1(t))]$ y $var[\log(\hat{s}_2(t))]$ se obtienen de (3.19) para los grupos A y B respectivamente.

De igual forma los límites del intervalo de confianza asimétrico sobre el rango de los riesgos acumulados se puede obtener usando la transformación complementaria log(log), por lo que se tiene:

$$\begin{aligned} \log[-\log \hat{s}_1(t)] - \log[-\log \hat{s}_2(t)] &= \log(\hat{H}_1(t)) - \log(\hat{H}_2(t)) \\ &= \log\Big(\frac{\hat{H}_1(t)}{\hat{H}_2(t)}\Big) \end{aligned}$$

$$\Rightarrow \quad var\Big[\log[-\log \hat{s}_1(t)] - \log[-\log \hat{s}_1(t)]\Big] = var\Big[\log[-\log \hat{s}_1(t)]\Big] + var\Big[\log[-\log \hat{s}_2(t)]\Big]$$

donde $var\Big[\log[-\log \hat{s}_1(t)]\Big]$ y $var\Big[\log[-\log \hat{s}_2(t)]\Big]$ corresponden a los grupos A y B respectivamente y son obtenidas de (3.20).

Sea t^* un tiempo específico; bajo la hipótesis nula

$$H_0 : s_1(t) = s_2(t) \quad vs \quad H_a : s_1(t) \neq s_2(t)$$

y sea $s^*(t)$ la función de supervivencia estimada desde los grupos combinados, la cual facilita un estimador constante. El cual se desarrolla a partir del estimador Kaplan-Meier, (3.13), con la probabilidad condicional de un evento estimado como:

$$\hat{Q}_i^* = \frac{d_{1i} + d_{2i}}{n_{1i} + n_{2i}}; \quad 1 \leq i \leq M$$

entonces para el grupo A y el grupo B a partir del uso de (3.8) se tiene la varianza estimada de $s_j(t^*)$ en el tiempo específico t^* bajo la hipótesis nula como:

$$var^*[\hat{s}_j(t^*)] = \left[\hat{s}(t^*)\right]^2 \left[\sum_{r=1;t_{(r)} \leq t^*}^{i} \frac{\hat{Q}_r^*}{n_{jr}(1 - \hat{Q}_r^*)} \right]; \quad j = 1, 2$$

donde n_{jr} es el número en riesgo para el grupo A y el grupo B, $j = 1$ y $j = 2$ respectivamente, en el r-ésimo tiempo del evento. Por lo que una prueba para muestras grandes bajo el supuesto de la igualdad de funciones de supervivencia en un tiempo t^* está dada por:

$$\chi^2 = \frac{([\hat{s}_1(t^*)] - [\hat{s}_2(t^*)])^2}{var([\hat{s}_1(t^*)]) + var([\hat{s}_1(t^*)])}$$

la cual se distribuye asintóticamente $\chi^2_{(1)}$.

3.5.1. Prueba de logrank

La prueba de logrank se usa para la comparación de dos grupos independientes, A y B, de datos de supervivencia. Primero se ordenan los tiempos de falla censurados y no. Supongase que se tienen r distintos tiempos de supervivencia, que además en el tiempo $t_{(i)}$, d_{1i} y d_{2i} elementos presentaron la falla en los grupos A y B respectivamente, $i = 1, 2, \ldots, r$, en caso de que dos o más elementos en un grupo presentan el mismo tiempo de falla, los valores de d_{1i} y d_{2i} serán 0 o 1. También supóngase que para el grupo A hay n_{1i} y que para B hay n_{2i} elementos en riesgo respectivamente justo antes de $t_{(i)}$, por lo que en el tiempo $t_{(i)}$ se tiene en total $d_i = d_{1i} + d_{2i}$ fallas y $n_i = n_{1i} + n_{2i}$ individuos en riesgo. Todo este desarrollo queda mejor esbozado en la tabla siguiente tabla.

	Número de fallas en $t_{(i)}$	Elemento que sobrevivió en $t_{(i)}$	Elemento en riesgo justo antes de $t_{(i)}$
Grupo A	d_{1i}	$n_{1i} - d_{1i}$	n_{1i}
Grupo B	d_{2i}	$n_{2i} - d_{2i}$	n_{2i}
Total	d_i	$n_i - d_i$	n_i

Tabla 3.3: Estructura de la tabla para la prueba logrank

Se busca probar la hipótesis nula:

$$H_0 : s_1(t) = s_2(t) \quad vs \quad H_a : s_1(t) \neq s_2(t)$$

es decir, que no hay diferencia entre las funciones de supervivencias de los dos grupos.

Considerándose que en la hipótesis nula la diferencia igual con cero es independiente del grupo, se toma como cierta y se tiene que las entradas de la tabla (3.3)quedan determinadas por el valor de d_{1i}, el número de fallas en $t_{(i)}$ del grupo A. Por lo que d_{1i} se establece como una variable aleatoria la cual toma valores desde 0 hasta el mínimo de d_i y n_{1i}, es decir, $d_{1i} \in [0, min(d_i, n_{1i})]$.

Observese que d_{1i} condicionada a d_i esta de hecho distribuida de forma hipergeométrica puesto que la probabilidad de que la variable aleatoria corresponda con el número de fallas en el grupo A es

$$\frac{\begin{pmatrix} d_i \\ d_{1i} \end{pmatrix} \begin{pmatrix} n_i - d_i \\ n_{1i} - d_{1i} \end{pmatrix}}{\begin{pmatrix} n_i \\ n_{1i} \end{pmatrix}}$$

donde $\begin{pmatrix} d_i \\ d_{1i} \end{pmatrix}$ representa el número de maneras en las que las fallas d_{1i} se pueden elegir desde las fallas d_i; $\begin{pmatrix} n_i - d_i \\ n_{1i} - d_{1i} \end{pmatrix}$ representa el número de maneras en las que el número de sobrevivientes $n_{1i} - d_{1i}$ se puede elegir desde $n_i - d_i$ y $\begin{pmatrix} n_i \\ n_{1i} \end{pmatrix}$ representa el número de maneras en las que el riesgo n_{1i} se puede elegir desde los riesgos n_i.

Recúerdese que $\begin{pmatrix} d_i \\ d_{1i} \end{pmatrix} = d_{1i}Cd_i = \frac{d_i!}{d_{1i}!(d_i - d_{1i})!}$

Por lo anterior d_{1i} tiene como valor medio esperado:

$$E(d_{1i}) = \frac{n_{1i}d_i}{n_i}$$

que se interpreta como el número de elementos que se espera que presenten la falla en el tiempo t_i para el grupo A. El valor de $E(d_{1i})$ bajo el supuesto de la hipótesis nula permite saber que la probabilidad de fallas en el tiempo $t_{(i)}$ no depende de si el elemento que vaya a presentar el evento pertenece al grupo A o al B, ya que la probabilidad de que se presente en $t_{(i)}$ es $\frac{d_i}{n_i}$.

Ahora se debe de considerar la información de la tabla de 2 por 2, es decir de (3.3), todas las medidas de las desviaciones de los valores observados de d_{1i} de sus valores esperados, se hace la suma de las diferencias $d_{1i} - E(d_{1i})$ del total de número de fallas del

grupo A y el grupo B, dando como resultado la estadística:

$$\Theta = \sum_{i=1}^{r} (d_{1i} - E(d_{1i}))$$

Nótese que esta estadística tiene media cero, además de ser la diferencia entre el total de fallas y el valor esperado total de fallas del grupo A.

Dado que los tiempos de falla son independientes, la varianza de Θ es la suma de las varianzas de d_{1i}. Ya que d_{1i} tiene una distribución hipergeométrica la varianza de d_{1i} es:

$$var(d_{1i}) = \frac{n_{1i} n_{2i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}$$

$$\Rightarrow \quad var(\Theta) = \sum_{i=i}^{r} var(d_{1i})$$

dado que $E(\Theta) = 0$ se tiene que

$$\frac{\Theta}{\sqrt{var(\Theta)}} \sim N(0,1) \tag{3.21}$$

Si (3.21) se eleva al cuadrado se obtiene una estadística Ω con una distribución Ji-cuadrada con un grado de libertad, es decir

$$\Omega = \frac{\Theta^2}{var(\Theta)} \sim \chi^2_{(1)}$$

Esta prueba estadística es importante por el hecho de que se maneja desde el rango de los tiempos de supervivencia de los grupos A y B. La Ω permite tener información de la magnitud para la cual los tiempos observados en los grupos de datos diverge desde que éstos se presentan bajo el supuesto de la hipótesis nula de que la supervivencia en ambos grupos no era diferente. Un valor grande de la estadística da una gran evidencia contra la hipótesis nula, debido a que la distribución de Ω es una Ji-cuadrada con un grado de libertad, P-valor (P-value) asociado con el resultado de la estadística se obtiene de las distribuciones de la función de una variable aleatoria Ji-cuadrada. El porcentaje de puntos de la Ji-cuadrada se usa para identificar el rango dentro del cual en P-valor es falso.

Capítulo 4

Modelo de riesgo proporcional

4.1. Modelo de regresión logística

Por lo general los modelos que más se estudian en regresión tienen la necesidad de que la distribución de la variable dependiente sea normal; incluso cuando ésta no es normal pero continua se busca aplicarle una transformación de tal forma que su función de probabilidad se distribuya como una gaussiana, además estás normalizaciones no son siempre adecuadas y entonces se recurre al empleo de modelos que comúnmente se les suele llamar modelos de regresión logística. Para emplearlo se retoma el echo de que la función de densidad de probabilidad para una variable Y con distribución binomial se define como:

$$f(Y) = \left(\begin{array}{c} n \\ Y \end{array} \right) p^Y (1-p)^{n-Y}$$

y para una variable Bernoulli se define como:

$$f(Y) = p^Y (1-p)^{1-Y}$$

Se sabe bien que en un proceso binomial puntual la probabilidad de éxito p y la de fracaso $q = 1 - p$, donde ambas están relacionadas por $p + q = 1$.

Para fines del modelo de regresión logística se usará el denominado cociente de proporcionalidad (odds) p/q el cual se conoce como un indicador de que tan probable es el éxito al fracaso.

El considerar un modelo de regresión que estudie una variable binomial permitirá saber si está depende o no de otras. Si se considera que una variable binomial con parámetro P es independiente de una variable X, entonces se cumple $P : p|X = x$, para todo valor observado x de X. Por lo que el modelo de regresión no da una función que permita saber que P está dependiendo de X determinando los cocientes que definan la relación de dependencia.

Debe de notarse que p sólo puede tomar valores de 0 a 1 y para p/q de 0 a ∞, mientras que para $\ln(p/q)$ quedan en todos los reales, $(-\infty, \infty)$. Entonces, se hace uso del cociente de proporcionalidad (odds) p/q y para una variable independiente X, el modelo de regresión logística está dado por la forma:

$$\ln\left(\frac{p}{q}\middle| X = x\right) = \beta_0 + \beta_1 X$$

o sencillamente como:

$$\ln(p/q) = \beta_0 + \beta_1 X \tag{4.1}$$

donde β_0 y β_1 son los parámetros de regresión y X es una variable que puede ser aleatoria o no, al igual que puede ser continua o discreta.

El modelo (4.1) se puede escribir de otras formas equivalentes que permitan en otros casos sus manejos más cómodamente.

$$\begin{aligned}
\ln(p/q) &= \beta_0 + \beta_1 X \\
\Rightarrow \ln\left(\frac{p}{1-p}\right) &= \beta_0 + \beta_1 X \\
\Rightarrow \frac{p}{1-p} &= e^{\beta_0+\beta_1 X} \\
\Rightarrow p &= (1-p)e^{\beta_0+\beta_1 X} \\
\Rightarrow p + pe^{\beta_0+\beta_1 X} &= e^{\beta_0+\beta_1 X}
\end{aligned}$$

$$\Rightarrow p = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{4.2}$$

y dividiendo la segunda parte de la última expresión por $e^{\beta_0+\beta_1 X}$ se tiene que

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}} \tag{4.3}$$

En las expresiones (4.2) y (2.14) se puede calcular el proceso binomial para los valores que toma la variable X. Ahora si se tiene $z = e^{\beta_0+\beta_1 X}$ entonces

$$f(z) = \frac{1}{1+e^{-z}}$$

la cual es conocida en otras muchas áreas de las matemáticas y se le denomina función logística.

El modelo de regresión logística también se puede generalizar para el caso de k variables independientes, quedando expresado de la siguiente forma:

$$\ln(p/q) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

de forma equivalente a (4.3)

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}}$$

Véase ahora una interpretación de los coeficientes para el modelo simple con una sola variable independiente. Supóngase, pero sólo para una interpretación de los coeficientes, que la variable X solo puede tomar valores 0 y 1. Para el valor $X = 0$ el modelo queda reducido a la expresión:

$$\ln(p/q|X = 0) = \beta_0 \tag{4.4}$$

de tal forma que β_0 es igual al logaritmo del cociente de proporcionalidad cuando la variable independiente es cero. Para cuando $X = 1$ y considerando (4.4) se tiene que:

$$\begin{aligned} \ln(p/q|X = 1) &= \beta_0 + \beta_1 \\ &= \ln(p/q|X = 0) + \beta_1 \\ \Rightarrow\ \beta_1 &= \ln(p/q|X = 1) - \ln(p/q|X = 0) \\ &= \ln\left(\frac{p/q|X = 1}{p/q|X = 0}\right) \end{aligned}$$

de tal forma que β_1 es el logaritmo de la razón de los cocientes de proporcionalidad (odds ratio, OR) para los valores de X. Si la variable binomial es independiente de X, entonces ambos cocientes de proporcionalidad son iguales, por lo que OR será igual a uno y su logaritmo sera cero. La ventaja de trabajar con un modelo de regresión logístico es que para mostrar la independencia de las variables basta con estudiar si el coeficiente β_1 es cero.

Otra medida que hay que destacar en la relación de una variable binomial y una independiente cualquiera es la llamada de riesgo relativo (RR) la cual resulta en muchos casos más fiable que el OR. Para el caso explicativo donde la variable independiente sólo puede tomar dos valores el riesgo relativo resulta ser

$$RR = \frac{p|X = 1}{p|X = 0}$$

considerando (4.3) se tiene que

$$RR = \frac{1 + e^{-\beta_0}}{1 + e^{-(\beta_0 + \beta_1)}}$$

Nótese que:

$$\begin{aligned} OR &= \frac{p/q|X = 1}{p/q|X = 0} \\ &= \frac{(p|X = 1)(q|X = 0)}{(p|X = 0)(q|X = 1)} \\ &= RR\left(\frac{q|X = 0}{q|X = 1)}\right) \end{aligned}$$

En esta situación si la variable binomial tiene una probabilidad muy pequeña el OR y RR tienen valores similares.

4.2. Método de Newton-Raphson

Considérese un sistema de dos ecuaciones con dos incógnitas el cual puede representarse por medio de dos funciones $f(x,y)$ y $g(x,y)$ cualesquiera en x e y de la siguiente forma:

$$f(x,y) = 0$$
$$g(x,y) = 0$$

Ahora se toma la aproximación de primer orden del desarrollo de la serie de Taylor alrededor del punto (x_0, y_0) donde x_0 y y_0 son valores escogidos de manera arbitraria. Para las funciones $f(x,y)$ y $g(x,y)$ se tiene:

$$0 = f(x,y) \approx f(x_0,y_0) + (x-x_0)\frac{df(x_0,y_0)}{dx} + (y-y_0)\frac{df(x_0,y_0)}{dy}$$

$$0 = g(x,y) \approx g(x_0,y_0) + (x-x_0)\frac{dg(x_0,y_0)}{dx} + (y-y_0)\frac{dg(x_0,y_0)}{dy}$$

si se considera la notación $\dfrac{dh(x,y)}{dx} = h'_x(x,y)$, entonces las dos ecuaciones anteriores se pueden reescribir de la siguiente forma:

$$f(x_0,y_0) + (x-x_0)f'_x(x_0,y_0) + (y-y_0)f'_y(x_0,y_0) \approx 0$$

$$g(x_0,y_0) + (x-x_0)g'_x(x_0,y_0) + (y-y_0)g'_y(x_0,y_0) \approx 0$$

$$\Rightarrow -f(x_0,y_0) = (x-x_0)f'_x(x_0,y_0) + (y-y_0)f'_y(x_0,y_0)$$
$$\Rightarrow -g(x_0,y_0) = (x-x_0)g'_x(x_0,y_0) + (y-y_0)g'_y(x_0,y_0)$$

Reescribiendo al sistema de ecuaciones en su forma matricial se tiene:

$$\begin{pmatrix} f'_x(x_0,y_0) & f'_y(x_0,y_0) \\ g'_x(x_0,y_0) & g'_x(x_0,y_0) \end{pmatrix} \begin{pmatrix} x-x_0 \\ y-y_0 \end{pmatrix} = -\begin{pmatrix} f(x_0,y_0) \\ g(x_0,y_0) \end{pmatrix}$$

si la matriz de derivadas parciales, la cual es conocida como el jacobiano, es invertible se tiene que:

$$\begin{pmatrix} x-x_0 \\ y-y_0 \end{pmatrix} = -\begin{pmatrix} f'_x(x_0,y_0) & f'_y(x_0,y_0) \\ g'_x(x_0,y_0) & g'_x(x_0,y_0) \end{pmatrix}^{-1} \begin{pmatrix} f(x_0,y_0) \\ g(x_0,y_0) \end{pmatrix}$$

por lo tanto la solución que se busca queda determinada por:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \begin{pmatrix} f'_x(x_0, y_0) & f'_y(x_0, y_0) \\ g'_x(x_0, y_0) & g'_x(x_0, y_0) \end{pmatrix}^{-1} \begin{pmatrix} f(x_0, y_0) \\ g(x_0, y_0) \end{pmatrix} \tag{4.5}$$

Se debe notar que la solución encontrada es sólo una aproximación, por lo que se debe de tomar nuevos valores para x_0 e y_0 y se realiza nuevamente el cálculo de (4.5) interactivamente o cíclicamente, las soluciones que se van encontrando son cada vez más cercanas a la solución real. El proceso deberá de terminar cuando se llegue al límite de convergencia, el cual se presenta cuando la diferencia entre las soluciones sucesivas es muy pequeña o cuando el valor de la solución encontrado sea demasiado pequeño.

El método de Newton-Raphson se puede generalizar para un sistema de k ecuaciones. Donde el resultado tiene una expresión equivalente a (4.5) las cuales se pueden representar por funciones cualesquiera de la siguiente manera:

$$f_i(x_1, x_2, \ldots, x_k) = 0 \qquad i = 1, 2, 3, \ldots, k$$

$$\Rightarrow \quad \overline{x} = \overline{x}_0 - J^{-1}F$$

donde los términos anteriores están dados por:

$$\overline{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix} \quad \overline{x}_0 = \begin{pmatrix} x_{1_0} \\ x_{2_0} \\ \vdots \\ x_{k_0} \end{pmatrix} \quad F = \begin{pmatrix} f_1(x_{1_0}, x_{2_0}, \ldots, x_{k_0}) \\ f_2(x_{1_0}, x_{2_0}, \ldots, x_{k_0}) \\ \vdots \\ f_k(x_{1_0}, x_{2_0}, \ldots, x_{k_0}) \end{pmatrix}$$

$$J = \begin{pmatrix} f'_{1x_1}(x_{1_0}, x_{2_0}, \ldots, x_{k_0}) & f'_{1x_2}(x_{1_0}, x_{2_0}, \ldots, x_{k_0}) & \cdots & f'_{1x_k}(x_{1_0}, x_{2_0}, \ldots, x_{k_0}) \\ f'_{2x_1}(x_{1_0}, x_{2_0}, \ldots, x_{k_0}) & f'_{2x_2}(x_{1_0}, x_{2_0}, \ldots, x_{k_0}) & \cdots & f'_{2x_k}(x_{1_0}, x_{2_0}, \ldots, x_{k_0}) \\ \vdots & \vdots & \vdots & \vdots \\ f'_{kx_1}(x_{1_0}, x_{2_0}, \ldots, x_{k_0}) & f'_{kx_2}(x_{1_0}, x_{2_0}, \ldots, x_{k_0}) & \cdots & f'_{kx_k}(x_{1_0}, x_{2_0}, \ldots, x_{k_0}) \end{pmatrix}$$

donde $f'_{ix_j}(x_{1_0}, x_{2_0}, \ldots, x_{k_0})$ es el valor de la derivada parcial de $f_i(x_{1_0}, x_{2_0}, \ldots, x_{k_0})$ con respecto a x_j evaluada en $(x_{1_0}, x_{2_0}, \ldots, x_{k_0})$.

4.3. Modelo de regresión de Cox

El modelo de riesgo proporcional mejor conocido como de Cox debido a que fue éste en 1972 quien lo expuso, el cual es empleado para el análisis de datos de supervivencia censurados para identificar los factores más convenientes para designar un proceso y un diagnóstico en un estudio. Éste tiene su mayor aplicación en las pruebas e investigaciones biomédicas. Algo que se tiene que notar es que al igual que los modelos de regresión logística las variables que se emplean en el de riesgo proporcional por lo general no siguen una distribución normal.

El modelo de Cox se centra en las situaciones donde el riesgo de falla en un tiempo específico T depende de los valores de $x_1, x_2, \ldots, x_k$ de k variables explicativas, se asume que el valor de éstas se registran en el tiempo de origen del estudio.

Para que se establezca la estructura del modelo de Cox se considera una muestra de tamaño n, el número de individuos en el estudio, y sea $x_i' = (x_{1i}, x_{2i}, \ldots, x_{ki})$ el vector de variables explicativas para cada individuo $i = 1, 2, \ldots, n$. Se tiene que el modelo de Cox representa el riesgo de falla (o muerte) para un individuo i, por lo que se establece como:

$$h(t, x_i) = h_0(t) exp(\underline{\beta}' \underline{x}_i) \qquad i = 1, 2, ..., n \tag{4.6}$$

otra forma equivalente es:

$$h(t, x_i) = h_0(t) exp\left(\sum_{i=1}^{k} \beta_i x_i\right) \tag{4.7}$$

donde $exp(\underline{\beta}' \underline{x}_i)$ es una función de los valores del vector de las variables explicativas para el individuo i, el vector $\underline{x}_i$ es conocido también como de covariantes, multiplicado por $\underline{\beta} = [\beta_1, \beta_2, \ldots, \beta_k]$ de coeficientes de regresión.

Nótese que la función $exp(\underline{\beta}' \underline{x}_i)$ no depende del tiempo, por lo que las variables explicativas se asumen constantes en éste. La función $h_0(t)$ es una función de riesgo la cual sólo depende del tiempo y se asume la misma para todos los individuos i, y es llamada de base y es por ésta que $exp(\underline{\beta}' \underline{x}_i)$ sea una función constante a lo largo del tiempo que y el modelo de Cox sea de riesgo proporcional. $h_0(t)$ puede verse también como la función para un individuo i, el cual tiene vector covariantes $\underline{x}_i = 0$, $h(t, 0) = h_0(t)$.

Por el hecho de que $h_0(t)$ no tiene una forma específica el modelo de Cox se considera no paramétrico, ya que en el caso práctico el tiempo suele considerarse como una variable discreta.

Otra forma de considerar el modelo de riesgo proporcional es en su expresión logarítmica, y es en ésta con la que se suele trabajar debido a la comodidad para estimar los coeficientes de regresión:

$$\ln\left(\frac{h(t, x_i)}{h_0(t)}\right) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{4.8}$$

Esta expresión permite observar el logaritmo del riesgo relativo, el cual se plantea en el modelo de regresión logística, por lo que (4.8) se puede bosquejar como un modelo de variables independientes (explicativas). Nótese que no existe un término constante β_0, si existiese lo que se hace es dividir la función de riesgo base $h_0(t)$ por $exp(\beta_0)$ y se construye una nueva $h_0^*(t) = h_0(t)/exp(\beta_0)$.

4.4. Estimación de los coeficientes de regresión del modelo de Cox

Para estimar los coeficientes de regresión del modelo de Cox, es necesario construir la función de verosimilitud, existen diversas maneras, pero es debido a Cox(1972) que toma como base la función de verosimilitud parcial (o relevante) para el modelo de riesgo proporcional por el hecho de que (4.6) no hace ningún supuesto sobre $h_0(t)$, la única información de los datos a la verosimilitud que se tienen es en los datos en los que se presentan eventos.

Supóngase que se consideran funciones de probabilidad continuas, por lo que no se puede presentar empates en los tiempos observados. Ahora se considera los datos de una muestra aleatoria de tamaño n formada por d tiempos de supervivencia observados ($d \leq n$; si se presentan eventos $d < n$, y si no existen pérdidas $d = n$) distintos $t_1 < t_2 < \cdots < t_d$ para los individuos $i = 1, 2, \ldots, d$ y para cada uno de ellos un vector $\underline{x}_i$ de observaciones de las covariantes (variables independientes), es decir,

$$\underline{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ki}) \qquad i = 1, 2, \ldots, d \text{ (incluidos para los que se presentan eventos)}$$

Para cada t_i existen n_i ($n_i = R(t_i)$ conjunto de individuos en riesgo en el tiempo t_i). Cox muestra que la verosimilitud de cada observación t_i está determinada por una probabilidad condicional, la cual es que en t_i ocurre un evento justamente para el individuo i condicionada a que existen n_i en riesgo:

$$\frac{h(t_i, \underline{x}_i)}{\sum_{l \in n_i} h(t_i, \underline{x}_l)} \tag{4.9}$$

es decir, el riego de un individuo i en el tiempo t_i dividido por la suma de los individuos l que están en riesgo en t_i. Si se considera la expresión del modelo (4.7), (4.9) se puede expresar como:

$$\frac{h_0(t_i) exp(\underline{\beta}' \underline{x}_i)}{\sum_{l \in n_i} h_0(t_i) exp(\underline{\beta}' \underline{x}_i)} = \frac{exp\left(\sum_{j=1}^{k} \beta_j x_{ji}\right)}{\sum_{l \in n_i} exp\left(\sum_{j=1}^{k} \beta_j x_{jl}\right)} \tag{4.10}$$

Por lo que la verosimilitud para toda la muestra, es el producto extendido a todas las observaciones:

$$L(\underline{\beta}) = \prod_{i=1}^{d} \frac{exp\left(\sum_{j=1}^{k} \beta_j x_{ji}\right)}{\sum_{l \in n_i} exp\left(\sum_{j=1}^{k} \beta_j x_{jl}\right)} \tag{4.11}$$

Si se considera a δ_i un indicador del evento, el cual toma el valor de cero si el individuo i, $i = 1, 2, \ldots, d$, es censurado por la derecha y uno en otro caso. La función de verosimilitud (4.11) queda expresada como:

$$L(\underline{\beta}) = \prod_{i=1}^{d} \left(\frac{exp\left(\sum_{j=1}^{k} \beta_j x_{ji} \right)}{\sum_{l \in n_i} exp\left(\sum_{j=1}^{k} \beta_j x_{jl} \right)} \right)^{\delta_i}$$

siendo x_{ji} la observación de la covariante (variable explicativa) j para el individuo i; la función de verosimilitud logarítmica correspondiente está dada por

$$\ln(L(\underline{\beta})) = \sum_{i=1}^{d} \delta_i \left(\sum_{j=1}^{k} \beta_j x_{ji} - \ln \left[\sum_{l \in n_i} exp\left(\sum_{j=1}^{k} \beta_j x_{jl} \right) \right] \right)$$

$$\Rightarrow \quad \ln(L(\underline{\beta})) = \sum_{i=1}^{d} \delta_i \sum_{j=1}^{k} \beta_j x_{ji} - \sum_{i=1}^{d} \delta_i \ln \left(\sum_{l \in n_i} exp\left(\sum_{j=1}^{k} \beta_j x_{jl} \right) \right)$$

Por lo que los estimadores para los coeficientes β_j serán aquellos que resulten de la solución del sistema de ecuaciones

$$\frac{\ln(L(\underline{\beta}))}{d\beta_j} = 0 \quad j = 1, 2, \ldots, k$$

el cual es no lineale, por lo que se resolverá haciendo uso del método de Newton-Raphson.

$$\frac{d \ln(L(\underline{\beta}))}{d\beta_j} = \sum_{i=1}^{d} \delta_i \left(x_{ji} - \frac{\sum_{l \in n_i} x_{jl} exp\left(\sum_{u=1}^{k} \beta_u x_{ul} \right)}{\sum_{l \in n_i} exp\left(\sum_{u=1}^{k} \beta_u x_{ul} \right)} \right) \qquad j = 1, 2, \ldots, k \tag{4.12}$$

$$\text{si} \quad A_{ji} = \frac{\sum_{l \in n_i} x_{jl} exp\left(\sum_{u=1}^{k} \beta_u x_{ul} \right)}{\sum_{l \in n_i} exp\left(\sum_{u=1}^{k} \beta_u x_{ul} \right)} = \frac{\sum_{l \in n_i} x_{jl} exp\left(\underline{\beta}' \underline{x}_l \right)}{\sum_{l \in n_i} exp\left(\underline{\beta}' \underline{x}_l \right)} \tag{4.13}$$

entonces (4.12) se puede escribir como:

$$\frac{\ln(L(\underline{\beta}))}{d\beta_j} = \sum_{i=1}^{d} \delta_i (x_{ji} - A_{ji}) \quad j = 1, 2, \ldots, k$$

Ahora se desarrolla el Jacobiano, que forma la matriz de información, cuya inversa es la de varianzas-covarianzas y que se compone de las segundas derivadas las cuales son:

$$\frac{d^2 \ln(L(\underline{\beta}))}{d\beta_u d\beta_v} = -\sum_{i=1}^{d} \delta_i \frac{\sum_{l \in n_i} x_{ul} x_{vl} exp\left(\sum_{j=1}^{k} \beta_j x_{jl}\right)}{\sum_{l \in n_i} exp\left(\sum_{j=1}^{k} \beta_j x_{jl}\right)} - A_{ui} A_{vi} \qquad u, v = 1, 2, \ldots, k \quad (4.14)$$

ahora la matriz de varianza-covarianzas se puede escribir como

$$v(\underline{\beta}) = cov(\underline{\beta}) = \left[-\frac{d^2 \ln[L(\underline{\beta})]}{d\underline{\beta}_u \underline{\beta}_v} \right] \quad (4.15)$$

el Jacobiano se representará por $J(\underline{\beta}_s)$ y $F = \left(\frac{d \ln(L(\underline{\beta}))}{d\beta_j}\right) = U(\underline{\beta}_s)$

Respecto al proceso de Newton-Raphson, en lo que concierne a la estimación del vector de coeficientes $\underline{\beta}$ en el ciclo $s+1$ del proceso de interacciones, o el paso $s+1$ de asignación de valores arbitrarios hasta encontrarse con el límite de convergencia, el cual se encuentra cuando el proceso de asignación termina y encuentra un valor para el cual la función verosimilitud (en este caso logarítmica) es muy pequeño o cuando la diferencia entre la estimación de los coeficientes sucesiva es muy pequeña, el vector $\hat{\underline{\beta}}_{s+1}$ es:

$$\hat{\underline{\beta}}_{s+1} = \hat{\underline{\beta}}_s + J^{-1}(\hat{\underline{\beta}}_s) U(\hat{\underline{\beta}}_s) \qquad s = 0, 1, 2, \ldots$$

donde $U(\hat{\underline{\beta}}_s)$ es el vector de asignación de valores eficientes y $J^{-1}(\hat{\underline{\beta}}_s)$ es la inversa de la matriz de información, ambos evaluados en $\hat{\underline{\beta}}_s$. El proceso puede comenzar en $s = 0$, $\hat{\underline{\beta}}_0 = \underline{0}$.

En el caso donde se considera al tiempo como una variable discreta es posible observar empates en los tiempos. Una versión discreta del modelo de riesgo proporcional (4.6) es:

$$\frac{h(t_i, \underline{x}_i)}{1 - h(t_i, \underline{x}_i)} = exp\left(\sum_{i=1}^{k} \beta_i x_i\right) \frac{h_0(t_i)}{1 - h_0(t_i)} \quad (4.16)$$

para la cual una función de verosimilitud fue sugerida por Cox(1972)

$$\prod_{j=1}^{m} \frac{exp\left(\sum_{i=1}^{k} \beta_i s_{ij}\right)}{\sum_{l \in R(t_j; d_j)} exp\left(\sum_{i=1}^{k} \beta_i s_{ij}\right)} \quad (4.17)$$

donde s_{ij} es el vector de sumas de las covariantes (variables independientes) de todos los individuos en los que los eventos ocurren en t_j, el tiempo de falla. Si el tamaño de la

muestra es n, que está formado por m tiempos, $t_1 < t_2 < \cdots < t_m$ en los cuales existen n_i individuos en riesgo y hay d_i eventos $(i = 1, 2, \ldots, m)$ y para cada evento una variable $\underline{x}_{ij}$ $i = 1, 2, \ldots, k,\ j = 1, 2, \ldots, m$ de covariantes. La notación $R(t_j; d_j)$ denota el conjunto de los d_j individuos empatados de los que están en riesgo n_j en el tiempo t_j.

El modelo (4.16) representa la probabilidad de que un individuo presente el evento en una unidad del intervalo de tiempo $(t, t + 1)$, condicionada a la supervivencia en un tiempo t. Si de hecho se toma el límite de la longitud del intervalo del tiempo discreto éste tiende a ser cero, entonces el modelo (4.16) tiende al (4.6). Cuando no hay empates, que es cuando $d_j = 1$ para cada tiempo de falla la aproximación de (4.17) se reduce a la forma de la ecuación (4.11).

4.5. Intervalos de confianza y prueba de hipótesis

Cuando se busca hallar el modelo de riesgo proporcional más eficaz, se encuentra la estimación de los parámetros, se tiene que considerar también a sus errores estándar (EE). Los errores estándar se usan para hallar un intervalo de confianza el cual permite ver un rango de que tan fiable es la estimación sobre los parámetros β desconocidos, con una confiabilidad del $(1 - \alpha)100\,\%$.

Por el teorema del límite central los estimadores obtenidos por máxima verosimilitud son asintóticamente normales. Y cuya matriz de varianzas-covarianzas es la inversa de la de información (Jacobiano), que permite obtener las varianzas de los parámetros β_i, que se ubican en la diagonal principal de esta matriz.

Ya que se conoce la varianza de los coeficientes β_i, $var(\hat{\beta}_i)$, un intervalo de confianza al $(1 - \alpha)100\,\%$ para el parámetro β_i es:

$$\hat{\beta}_i \pm Z_{1-\alpha/2} EE(\hat{\beta}_i)$$

donde $Z_{1-\alpha/2}$ es el cuantil de una normal estándar. Sin embargo hay que resaltar que los estimadores obtenidos no son los usuales de una regresión lineal, sino que representan el cociente de riesgo (o el conocido en regresión logística como odds ratio o cociente de proporcionalidad), por lo que los intervalos de confianza que se están buscando son:

$$\exp^{\hat{\beta}_i \pm Z_{\alpha/2} EE(\hat{\beta}_i)}$$

Se busca probar la hipótesis nula $H_0 : \beta_i \ \ vs \ \ H_a : \beta_i \neq 0$.

Se considera el estadístico $Z = \dfrac{\hat{\beta}_i - \beta_i}{EE(\hat{\beta}_i)} \ \sim \ N(0, 1)$
donde a es una constante y la región crítica es $|Z| > Z_{\alpha/2}$, elevando al cuadrado a Z se tiene el estadístico:

$$W = \frac{(\hat{\beta}_i - a)^2}{var(\hat{\beta}_i)}$$

que se distribuye como una Ji-cuadrada con un grado de libertad, por lo que la región crítica para el contraste es $W > \chi^2_{(\alpha)}$. Al estadístico W se le conoce como estadística o prueba de Wald.

La hipótesis nula importante a contrastar es para $\beta_i = 0 \;\; i = 1, 2, \ldots, k$. No rechazar esta hipótesis para algún valor i, muestra que la variable $\underline{x}$ no depende de x_i y por lo tanto el parámetro β_i asociado a x_i no se debe de considerar en el modelo, es decir, el término $\beta_i x_i$ no debe de aparecer en el modelo, pues no da información suficiente o necesaria para el pronóstico que se busca hacer.

4.6. Estimación de la función de supervivencia y de riesgo base

En esta sección se busca tener una expresión de las funciones de supervivencia y la de riesgo base estimadas.

Se tiene el modelo de riesgo proporcional

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

$$\Rightarrow \int_0^t h(u, \underline{x}) du = \int_0^t h_0(u) exp(\underline{\beta}' \underline{x}) du$$

$$\Rightarrow H(t, \underline{x}) = H_0(t) exp(\underline{\beta}' \underline{x})$$

donde $H(t, \underline{x})$ y $H(t)_0$ representan el riesgo y el riesgo base acumulados respectivamente.

Nótese que para el caso discreto $\left(\text{tomando } \sum_{i=0}^{t}\right)$ se tiene una expresión análoga.

Ahora se toma $-H(t, \underline{x}) = -H_0(t) exp(\underline{\beta}' \underline{x})$

$$\begin{aligned} \Rightarrow \exp[-H(t, \underline{x})] &= \exp[-H_0(t) exp(\underline{\beta}' \underline{x})] \\ &= [\exp(-H_0(t))]^{exp(\underline{\beta}' \underline{x})} \end{aligned}$$

$$\Rightarrow \quad s(t, \underline{x}) = [s_0(t)]^{exp(\underline{\beta}' \underline{x})} \tag{4.18}$$

Por lo que una función de supervivencia de una variable T con vector de covariantes $\underline{x}$, donde $s_0(t)$ se considera la función de supervivencia arbitraria (discreta, continua o mixta) queda expresada en (4.18).

Estimar $h_0(t)$ es equivalente a estimar $s_0(t)$. Para que se realice la estimación de $s_0(t)$ se considera la función de verosimilitud parcial y el proceso es análogo al que se uso para obtener el estimador Kaplan-Meier.

Sean $t_{(1)}, t_{(2)}, \ldots, t_{(k)}$ los tiempos de falla distintos, sea D_i el conjunto que representa al número de individuos que presentan la falla en $t_{(i)}$ y sea C_i el conjunto que representa a número de individuos censurados en el intervalo $[t_i, t_{i+1}] \quad i = 1, 2, \ldots, k$ donde $t_0 = 0$ y $t_{k+1} = \infty$. Los tiempos censurados en el intervalo $[t_i, t_{i+1}]$ se representan por t_l donde l recorre a el número de individuos censurados.

La información que aporta un individuo, con covariantes $\underline{x}$ que falla en $t_{(i)}$, a la verosimilitud bajo la independencia de censura es

$$s_0(t_{(i)})^{exp(\underline{\beta}'\underline{x})} - s_0(t_{(i)} + 0)^{exp(\underline{\beta}'\underline{x})}$$

y la información de una observación censurada está dada por

$$s_0(t + 0)^{exp(\underline{\beta}'\underline{x})}$$

entonces la función de verosimilitud se determina como

$$L = \prod_{i=0}^{k} \left[\prod_{l \in D_i} \left(s_0(t_{(i)})^{exp(\underline{\beta}'\underline{x}_l)} - s_0(t_{(i)} + 0)^{exp(\underline{\beta}'\underline{x}_l)} \right) \prod_{l \in C_i} s_0(t + 0)^{exp(\underline{\beta}'\underline{x}_l)} \right]$$

donde D_0 es vacío.

Para maximizar L se toma $s_0(t) = s_0(t_{(i)}+0)$ para $t_{(i)} < t \leq t_{(i+1)}$ y una probabilidad de tal forma que se presenten los tiempos de falla observados $t_{(1)}, t_{(2)}, \ldots, t_{(k)}$.

Estas consideraciones hacen que se considere un modelo discreto con una contribución de riesgo $1 - \varsigma_i$ en $t_{(i)}$ $(i = 1, 2, \ldots, k)$. Por lo que se toma

$$s_0(t_{(i)}) = s_0(t_{(i-1)} + 0) = \prod_{i=0}^{i-1} \varsigma_i \quad i = 1, 2, \ldots, k$$

donde $\varsigma_0 = 1$. Ahora sustituyendo en (4.18) y reorganizado se tiene

$$L = \prod_{i=0}^{k} \left[\prod_{j \in D_i} \left(1 - \varsigma_i^{exp(\underline{\beta}'\underline{x}_j)} \right) \prod_{l \in R(t_{(i)}) - D_i} \varsigma_i^{exp(\underline{\beta}'\underline{x}_l)} \right] \tag{4.19}$$

la función de verosimilitud a maximizar, donde $R(t_{(i)})$ es el conjunto de individuos en riesgo en $t_{(i)}$. A partir de aquí se considerará a $\underline{\beta} = \underline{\hat{\beta}}$ como el estimador obtenido en la sección anterior y (4.19) se maximiza respecto a los ς_i. Se toma ahora el logaritmo de L

$$\log(L) = \sum_{i=1}^{k} \left(\sum_{j \in D_i} \log \left(1 - \varsigma_i^{exp(\underline{\hat{\beta}}'\underline{x}_j)} \right) \right) + \sum_{l \in R(t_{(i)}) - D_i} exp(\underline{\hat{\beta}}'\underline{x}_l) \log(\varsigma_i)$$

Se toma la primera derivada igualada a cero de $\log(L)$ con respecto a ς_i, que da el estimador de máxima verosimilitud de ς_i como una solución para

$$\sum_{j \in D_i} \frac{exp(\underline{\hat{\beta}}'\underline{x}_j)}{1 - \hat{\varsigma}_i^{exp(\underline{\hat{\beta}}'\underline{x}_j)}} = \sum_{l \in R(t_{(i)})} exp(\underline{\hat{\beta}}'\underline{x}_l) \tag{4.20}$$

Si una falla ocurre en $t_{(i)}$ de manera individual, (4.20) puede ser resuelta para $\hat{\varsigma}_i$ directamente por

$$\hat{\varsigma}_i = \left(1 - \frac{exp(\underline{\hat{\beta}}' \underline{x}_{(i)})}{\sum_{l \in R(t_{(i)})} exp(\underline{\hat{\beta}}' \underline{x}_l)} \right)^{exp(-\underline{\hat{\beta}}' \underline{x}_{(i)})}$$

Por lo tanto la función de supervivencia base estimada por máxima verosimilitud es entonces:

$$s_0 = \prod_{i|t_{(i)}<t} \hat{\varsigma}_i \tag{4.21}$$

por lo que la función de supervivencia para un vector de covariantes $\underline{x}$ estimada es

$$s(t, \underline{x}) = \prod_{i|t_{(i)}<t} \hat{\varsigma}_i^{exp(\underline{\hat{\beta}}' \underline{x})}$$

En el caso donde existen tiempos de supervivencia empatados la función de supervivencia base, s_0, estimada requiere de una solución dado por un proceso interactivo. Este proceso interactivo se encuentra aproximando la suma sobre el lado izquierdo de la ecuación (4.20). Ahora reescribiendo

$$\hat{\varsigma}_i^{exp(\underline{\hat{\beta}}' \underline{x}_j)} = e^{exp(\underline{\hat{\beta}}' \underline{x}_j) \log(\hat{\varsigma}_i)}$$

se toma los dos primeros términos dados por la expansión de los exponentes

$$e^{exp(\underline{\hat{\beta}}' \underline{x}_j) \log(\hat{\varsigma}_i)} \approx 1 + exp(\underline{\hat{\beta}}' \underline{x}_j) \log(\hat{\varsigma}_i)$$

se sustituye la aproximación en (4.20)

$$\begin{aligned}
\sum_{j \in D_i} \frac{exp(\underline{\hat{\beta}}' \underline{x}_j)}{1 - \left[1 + exp(\underline{\hat{\beta}}' \underline{x}_j) \log(\hat{\varsigma}_i) \right]} &= \sum_{l \in R(t_{(i)})} exp(\underline{\hat{\beta}}' \underline{x}_l) \\
\Rightarrow \quad - \sum_{j \in D_i} \frac{1}{\log(\hat{\varsigma}_i)} &= \sum_{l \in R(t_{(i)})} exp(\underline{\hat{\beta}}' \underline{x}_l) \\
\Rightarrow \quad - \frac{d_i}{\log(\hat{\varsigma}_i)} &= \sum_{l \in R(t_{(i)})} exp(\underline{\hat{\beta}}' \underline{x}_l) \\
\Rightarrow \quad \hat{\varsigma}_i &= \exp \left(\frac{-d_i}{\sum_{l \in R(t_{(i)})} exp(\underline{\hat{\beta}}' \underline{x}_l)} \right)
\end{aligned}$$

donde d_i es el número de fallas en $t_{(i)}$, además la función de supervivencia base estimada para el caso en que existen tiempos de falla empatados está dada por

$$\hat{s}_0(t) = \prod_{i|t_{(i)}<t} \exp\left(\frac{-d_i}{\sum_{l\in R(t_{(i)})} exp(\hat{\underline{\beta}}'\underline{x}_l)}\right) \tag{4.22}$$

La función de riesgo base acumulada, H_0, estimada se obtiene de la relación que hay con $\hat{s}_0(t)$ y es

$$\hat{H}_0(t) = -\log(\hat{s}_0(t)) = \sum_{i|t_{(i)}<t} \frac{d_i}{\sum_{l\in R(t_{(i)})} exp(\hat{\underline{\beta}}'\underline{x}_l)} \tag{4.23}$$

Este estimador es conocido como el estimador de Breslow de la función de riesgo acumulada base.

En el caso donde el vector de covariantes es cero, $\underline{x} = 0$, y los tiempos de falla no son empatados se tiene que la ecuación (4.20) se establece como

$$\frac{d_i}{1-\hat{\varsigma}_i} = n_i \quad \Rightarrow \quad \hat{\varsigma}_i = 1 - \frac{d_i}{n_i}$$

entonces el estimador de la función de riesgo base es

$$\hat{h}_0(t) = \frac{d_i}{n_i} = 1 - \hat{\varsigma}_i$$

por lo que si se considera la función de supervivencia base de (4.21) se tiene que

$$\hat{s}_0(t) = \prod_{i=1}^{k}\left(1 - \frac{d_i}{n_i}\right)$$

el cual es el conocido estimador Kaplan-Meier.

Si se considera la ecuación (4.22) se tiene que

$$\hat{s}_0(t) = \prod_{i=1}^{k}\exp\left(-\frac{d_i}{n_i}\right)$$

esta función de supervivencia estimada es conocida como el estimador Nelson-Alen de la función de supervivencia, y al igual que para (4.23) se tiene que

$$\hat{H}_0(t) = \sum_{i=1}^{k}\frac{d_i}{n_i}$$

el cual es el estimador de Nelson-Alen dado en (3.17).

4.7. Estratificación del modelo de Cox

Una extensión del modelo de Cox es que permita múltiples estratos. El estrato separa a los elementos dentro de grupos disjuntos, cada grupo tiene un función de riesgo base distinta. Visto de otra forma, es que si se tienen dos vectores distintos de covariantes, $\underline{x}_1$ y $\underline{x}_2$, sus funciones de riesgo, o modelos de riesgo proporcional, estén relacionadas.

Supóngase que hay un factor que ocurre sobre m estratos, se define el modelo de riesgo proporcional para el individuo en i-ésimo estrato de este factor como

$$h_i(t, \underline{x}) = h_{0i} \exp(\underline{\beta}' \underline{x}) \quad i = 1, 2, \ldots, m \tag{4.24}$$

donde $\underline{\mathrm{x}}$ es el vector de covariantes. Las funciones de riesgo, $h_{01}, h_{02}, \ldots, h_{0m}$, para el m estrato son funciones arbitrarias del tiempo de falla y no están relacionada en su totalidad.

El modelo(4.24) asume que los individuos en el r-ésimo estrato los cuales pueden tener diferentes covariantes $\underline{x}_1$ y $\underline{x}_2$ aun tienen un cociente de proporcionalidad

$$\frac{h_r(t, \underline{x}_1)}{h_r(t, \underline{x}_2)} = \exp[\underline{\beta}'(\underline{x}_1 - \underline{x}_2)]$$

Nótese que los parámetros de regresión $\underline{\beta}$ no dependen del estrato, entonces el efecto de las covariantes es asumido como el mismo para todos los estratos, es decir, es común el valor del vector $\underline{\beta}$ ya que de lo contrario los estratos podrían considerarse conjuntos de datos distintos y se realizaría un análisis separadamente.

Se considera el cálculo de censura por la derecha y los empates en el conjunto de datos, para que se estimen los coeficientes de regresión usando una función de verosimilitud parcial, la cual describe una situación más general que en (4.11) y se construye como

$$L(\underline{\beta}) = \prod_{r=1}^{m} L_r(\underline{\beta}) \tag{4.25}$$

donde $L_r(\underline{\beta})$ es la función de verosimilitud de $\underline{\beta}$, dada en (4.11), y comienza sólo desde r-ésimo el estrato. La estimación de los parámetros $\underline{\beta}$, o la maximización de (4.25), y se realiza haciendo uso del método de Newton-Raphson. La pérdida de eficiencia que se presenta en la estimación del vector $\underline{\beta}$ cuando se hace uso de la estratificación innecesaria es poca (Kalblfleisch and Pretince, 1980, p.112).

La prueba de hipótesis que se usa sobre el vector de parámetros es $\beta_i = 0$ es la prueba usual que se realiza sobre las propiedades de la función de verosimilitud.

4.7.1. Variables dependientes del tiempo

Se ha mostrado la dependencia de la función de riesgo de una variable explicativa. Ahora se mostrará que existen variables explicativas que son dependientes del tiempo.

A las variables cuyos valores se ven afectados cada vez que el tiempo cambia en el estudio (investigación, proceso, etc.) se les conoce como variables dependientes del tiempo, es decir, que para el tiempo t la función de riesgo para el individuo i depende del valor de la variable explicativa $x_i(t)$ en el tiempo t y para el tiempo t^* depende del vector de la variable explicativa $x_i(t^*)$ en el tiempo t^* y así sucesivamente.

Las variables dependientes del tiempo están divididas en dos amplias categorías de variables dependientes del tiempo, las cuales son llamadas variables externas y variables internas.

Variables externas

Los valores que una variable externa (covariante externa) toma en un tiempo son ajenos en su totalidad al estudio de un individuo i, es decir, estos valores no están influenciados por el estudio del individuo. Los valores son generados por un llamado mecanismo externo para el individuo, por ejemplo la edad de un individuo que un estudio corto no afecta los resultados, sin embargo cuando el estudio es demasiado extenso la edad juega un factor muy importante en el individuo. Existe otro tipo de variable externa llamada *Auxiliar*, la cual presenta valores de un proceso estocástico externo para el individuo en estudio, por ejemplo una variable externa que mida la contaminación aérea como un factor que predice las frecuencias de ataques de asma.

Variables internas

A diferencia de las variables externas los valores de las variables internas resultan de un proceso estocástico generado por el individuo en estudio, es decir, las variables explicativas internas $x_i(t)$ en cada tiempo t es determinada por la experiencia del individuo i en el estudio, por ejemplo como responden los pacientes a un tratamiento de un tumor es registrado y supervisado.

Por lo tanto, el modelo de regresión de Cox y el modelo de Cox estratificado en términos de una variable explicativa dependiente del tiempo se puede representar como:

$$h_i(t, x_i(t)) = h_i(t) \exp(\underline{\beta}' \underline{x}(t))$$

$$h_i(t, x_i(t)) = h_{0i}(t) \exp(\underline{\beta}' \underline{x}(t))$$

4.8. Residuales

Cuando un modelo se ha ajustado para la representación de datos de supervivencia, se necesita saber si existe una covariante influenciada o si esta covariante no da información que no podría ser importante para el estudio. La elección de un modelo de Cox adecuado requiere de la revisión de varios procesos los cuales están basados en los cuantiles, que

en el campo de la regresión se conocen mejor como residuales, los cuales son importantes para la identificación de que tiempos de supervivencia tienen mayor valor o sirven para detectar que covariante de individuo en el estudio puede ser de baja o alta importancia en el modelo que se esta buscando ajustar.

Existen diversos residuales que han sido propuestos para el modelo de Cox, debido que los que se usan habitualmente en los modelos de regresión lineal no son adecuados para buscar una respuesta al ajuste del modelo de Cox. Para estos residuales se considerará los tiempos de supervivencia adecuados de n individuos, r tiempos de falla y $n - r$ serán los tiempos de falla censurados por la derecha.

Considérese también el modelo:

$$\hat{h}_i(t, \underline{x}) = \exp(\underline{\hat{\beta}}' \underline{x}) \hat{h}_0(t)$$

donde $\underline{\hat{\beta}}$ es un vector de covariantes estimado de tamaño (px1) y $\hat{h}_0(t)$ es la función de riesgo base estimada.

4.8.1. Residual Cox-Snell

Dentro del conjunto de residuales para el modelo de Cox, se tiene el residual Cox-Snell (dado por Cox y Snell 1968) el cual es de los más usados para el análisis de ajuste de modelo de Cox, respecto al análisis de datos de supervivencia. Además de una de las características de este residual es que puede ser aplicado a cualquier modelo paramétrico.

El residual de Cox-Snell para el i-ésimo individuo con tiempo de supervivencia t_i, puede ser censurado o no censurado, de define como

$$r_{csi} = -\log\left[\hat{s}(t_i, x_i)\right] = \exp(\underline{\hat{\beta}}' \underline{x}) \hat{H}_0(t) \quad i = 1, 2, \ldots, n \tag{4.26}$$

donde $\hat{s}(t_i)$ es la función de supervivencia estimada por Kaplan-Meier.

Note que el valor de r_{csi} es el valor del riesgo acumulado $\hat{H}(t_i, x_i) = -\log\left[\hat{s}(t_i, x_i)\right]$, además nótese también que si t_i es censurado entonces r_{csi} también es censurado.

Si se considera a r_{csi} como una variable aleatoria ésta sigue una distribución exponencial con media uno, esto se puede saber dado el siguiente resultado.

$$\text{Si } \; r_{cs} = -\log\left[s(t)\right] \; \Rightarrow \; t = s^{-1}\left[\exp(-r_{cs})\right] = s^{-1}(e^{-r_{cs}})$$

entonces la función de densidad de r_{csi} es $f_{r_{cs}}(r) = e^{-r}$, debido a que

$$\begin{aligned} f_{r_{cs}}(r) &= \frac{f_T\left[s^{-1}(e^{-r})\right]}{\left|\frac{dr}{dt}\right|} \\ &= \frac{f_T\left[s^{-1}(e^{-r})\right]}{\left|\frac{d(-\log[s(t)])}{dt}\right|} \\ &= \frac{f_T\left[s^{-1}(e^{-r})\right]}{\left|\frac{f_t(t)}{s(t)}\right|} \\ &= \frac{f_T\left[s^{-1}(e^{-r})\right]}{\left|\frac{f_t\left[s^{-1}(e^{-r})\right]}{s\left[s^{-1}(e^{-r})\right]}\right|} \\ &\therefore\ f_{r_{cs}}(r) = e^{-r} \end{aligned}$$

donde $f_t(t)$ es la función de densidad de probabilidad de la variable T que puede tomar valores de $t = t_i$ y r es la variable aleatoria que puede tomar valores de $r = r_{csi};\ i = 1, \ldots, n$.

Si se tiene que un individuo es censurado por la derecha, es decir, en el tiempo t_i^+ y el modelo de Cox ajustado es correcto se tendrá que el valor del residual en t_i+ tendrá el valor más pequeño de todas las observaciones no censuradas en el tiempo t_i. Dada esta observación se propusieron dos modificaciones al residual de Cox-Snell por Crowley y Hu (1977).

A los residuales modificados de Cox-Snell se les conoce como residuales de exceso y el primero está dado por

$$r'_{csi} = \begin{cases} r_{csi} & \text{si la observación es no censurada} \\ r_{csi} + 1 & \text{si la observación es censurada} \end{cases}$$

este hecho se da dado que r_{csi} tiene una distribución exponencial unitaria por lo que el primer residual modificado r'_{csi} tiene también una distribución exponencial unitaria, por lo que el valor excedido es uno.

Mientras que el primer residual modificado de Cox-Snell se considera la media uno, para el segundo residual modificado Crowley Hu utilizan la mediana para una distribución unitaria cuya función de supervivencia está dada por e^{-t}, por lo que la mediana es $e^{-t(50)} = 0.5$, entonces se tiene que

$$t(50) = \log(2) = 0.693$$

y el segundo residual de Cox-Snell está dado por

$$r''_{csi} = \begin{cases} r_{csi} & \text{si la observación es no censurada} \\ r_{csi} + 0.693 & \text{si la observación es censurada} \end{cases}$$

La manera de operar el uso de los residuales de Cox-Snell se puede hacer de la siguiente forma:

1. Se calcula el residual $r_{csi} = -\log\left[\hat{s}(t_i)\right]$ donde $\hat{s}(t_i)$ esta estimada por Kaplan-Meier.
2. Se calcula $\hat{s}_r(r_{csi})$ y se estima usando el mismo procedimiento que se usa para obtener el estimador Kaplan-Meier, y se calcula $-\log\left[-\hat{s}_r(r_{csi})\right],\ \ i = 1, \ldots, n.$
3. Se grafica r_{csi} contra $-\log\left[-\hat{s}_r(r_{csi})\right]$, si la gráfica de los datos $(r_{csi}, -\log\left[-\hat{s}_r(r_{csi})\right])$ es descrita por una línea recta con pendiente uno y término constante cero, entonces el modelo se puede considerar apropiado.

4.8.2. Residual martingala

Otro residual propuesto para ajustar un modelo de riesgo proporcional y que tenga propiedades similares a las del análisis de regresión lineal, es el residual conocido como residual Martingala, el cual se define como

$$r_{mi} = \delta_i - r_{csi} \tag{4.27}$$

donde δ_i es un indicador que toma valores de cero para el caso en que las observaciones del tiempo de supervivencia del i-ésimo individuo son censuradas y uno si es no censurada, además de que los valores de r_{mi} están dentro del intervalo $(-\infty, 1)$, y los residuales toman valores negativos cuando $\delta_i = 0$, es decir, cuando la observación es censurada. Los residuales Martingala tienen una distribución sesgada con media cero.

4.8.3. Residual de desviación

Therneau y Grambsch (1990) introdujeron los residuales de desviación, los cuales a diferencia de los residuales Martingala tienen una mayor distribución simétrica alrededor del cero, y los cuales están definidos como

$$r_{di} = sign(r_{mi})\sqrt{-2\left[r_{mi} - \delta_i \log(\delta_i - r_{mi})\right]} \tag{4.28}$$

donde $sign(*)$ es la función signo la cual se define como:

$$sign(r_{mi}) = \begin{cases} 1 & \text{si } \ r_{mi} > 0 \\ 0 & \text{si } \ r_{mi} = 0 \\ -1 & \text{si } \ r_{mi} < 0 \end{cases}$$

de aquí que los residuales r_{di} tiene el mismo signo que los residuales Martingala.

Nótese que los residuales de desviación son una transformación de los residuales Martingala para que los valores obtenidos tengan una distribución simétrica alrededor del cero para cuando el modelo es apropiado.

4.8.4. Residual de Schoenfeld

A diferencia de los tres residuales mencionados con anterioridad, los cuales se basan fundamentalmente en la función de riesgo acumulada, $\hat{H}(t, \underline{x})$, Schoenfeld (1982) propuso un residual el cual se basa para cada persona y cada covariante, además de estar basado en la primera derivada de la función de verosimilitud logarítmica, descrita en (4.12). Se define el residual de Schoenfeld para la j-ésima covariante del i-ésimo individuo en el tiempo de supervivencia t_i como

$$r_{sji} = \delta_i \left[x_{ji} - \frac{\sum_{l \in n_i} x_{jl} \exp\left(\underline{\hat{\beta}}' x_l\right)}{\sum_{l \in n_i} \exp\left(\underline{\hat{\beta}}' x_l\right)} \right] \qquad i = 1, \ldots, n, \ j = 1, \ldots, p \tag{4.29}$$

donde δ_i es el indicador ya conocido y x_{ji} es el valor de la covariante j-ésima para el i-ésimo individuo en el estudio, n_i es el conjunto de individuos en riesgo y $\underline{\hat{\beta}}$ es el vector de coeficientes estimado.

Debe de notarse que los valores diferentes de cero sólo existen cuando los tiempos de supervivencia no son censurados. Dado que los residuales de Schoenfeld están basados en (4.12) su suma es cero por lo que asintóticamente tiene una media de valor cero por lo que no está relacionado con algún otro.

Grambsh y Therneau (1994) propusieron una modificación a los residuales de Schoenfeld los cuales son influenciados por la matriz de varianzas-covarianzas del vector $\underline{r_{si}} = (r_{s1i}, r_{s2i}, \ldots, r_{spi})$ es más preciso al determinar la desviación en el modelo de Cox sugerido y se define como

$$r*_{si} = rvar(\underline{\hat{\beta}})\underline{r_{si}}$$

donde r es el número de tiempos de supervivencias observados no censurados o el número de fallas y $var(\underline{\hat{\beta}})$ es la matriz de varianzas-covarianzas del vector $\underline{\hat{\beta}}$ descrito en (4.15).

Los residuales de desviación y de Schoenfeld se gráfica contra el tiempo de supervivencia o una covariante, tal gráfica se puede usar para la detección de ciertos patrones los cuales sirven para detectar la existencia de puntos aberrantes (outliers) o problemas de estabilidad del modelo.

4.8.5. Residual puntual (Score)

El residual puntual se considera una modificación del residual de Schoenfeld, dado que también se define a partir de la primera derivada de la función de verosimilitud parcial logarítmica respecto a las covariantes, β_j, de la cual se obtiene el valor adecuado (o puntual) para las β_j, y se define como

$$r_{pji} = \delta_i(x_{ji} - \hat{A}_{ji}) + \exp\left(\underline{\hat{\beta}}\underline{x}_i\right) \sum_{t_r \leq t_i} \frac{(\hat{A}_{jr} - x_{ji})\delta_r}{\sum_{l \in n_r} \exp\left(\underline{\hat{\beta}}' \underline{x}_l\right)}$$

donde x_{ji} es el valor del i-ésimo individuo de la j-ésima covariante, δ_i y δ_r es el indicador ya conocido, n_r es el conjunto de individuos en riesgo en el tiempo de supervivencia t_r y

$$\hat{A}_{ji} = \frac{\sum_{l \in n_i} x_{jl} \exp\left(\hat{\underline{\beta}}' \underline{x}_l\right)}{\sum_{l \in n_i} \exp\left(\hat{\underline{\beta}}' \underline{x}_l\right)}$$

de aquí que la información de la i-ésima observación para la derivada depende sólo de la información hasta el tiempo t_i.

Capítulo 5

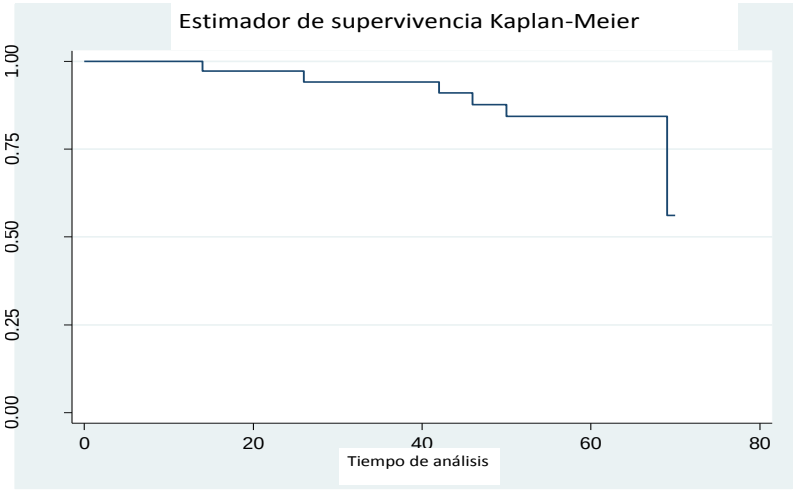
Ejemplo 1. Comparación de dos tratamiento de cáncer de próstata

En este capítulo y el siguiente se abordará el manejo y análisis de datos de supervivencia. Dichos ejemplos son más una motivación al uso de la teoría descrita en este trabajo, que el de buscar un resultado novedoso.

La base de datos para este ejemplo fue tomada de un ensayo clínico para comparar tratamientos para el cáncer de próstata, al tratamiento que se le asigna valor uno es un placebo y el de valor dos es *dietilestilbestrol* (conocido también como DES, el cual es un estrógeno sintético utilizado durante años para disminuir el riesgo de aborto en mujeres embarazadas y para tratar problemas de próstata). Además de tener en cuenta el tiempo de supervivencia medido en meses en el estudio de los pacientes en el ensayo clínico se consideraron otros datos, tales como el nivel de hemoglobina medido en gr/100ml, el tamaño del tumor primario medido en cm^2, la edad del paciente y el valor de un índice que combina el tamaño y estado del tumor, a este se le conoce como de Gleason. Lo que determino fue si los pacientes que son tratados con el dietilestilbestrol viven más que aquellos que son tratados con un placebo. La tabla (5.1) muestra los datos correspondientes a los 38 pacientes sobre los cuales se realizó el ensayo clínico.

Lo primero que se calcula es la función de supervivencia estimada por Kaplan-Meier para conocer la probabilidad acumulada de cada paciente a lo largo del tiempo. Esta probabilidad se describe en la columna 5 de la tabla (5.2), ésta tabla se obtuvo como resultado del paquete STATA, y se describe su comportamiento en la gráfica (5.1), la cual se obtuvo del paquete R; también se obtiene la tabla de vida, tabla (5.3) la cual se obtuvo con STATA y su gráfica, la cual coincide con la gráfica (5.2) que se encontró con R, que de igual manera se refleja la supervivencia de los pacientes en periodos de 5 meses; en esta tabla se puede expresar que los valores de la probabilidad de supervivencia para los pacientes de cáncer de próstata es más fiable, por el hecho de que los errores de estimación son más pequeños.

Después se probó la hipótesis nula de que las funciones de supervivencia son iguales,

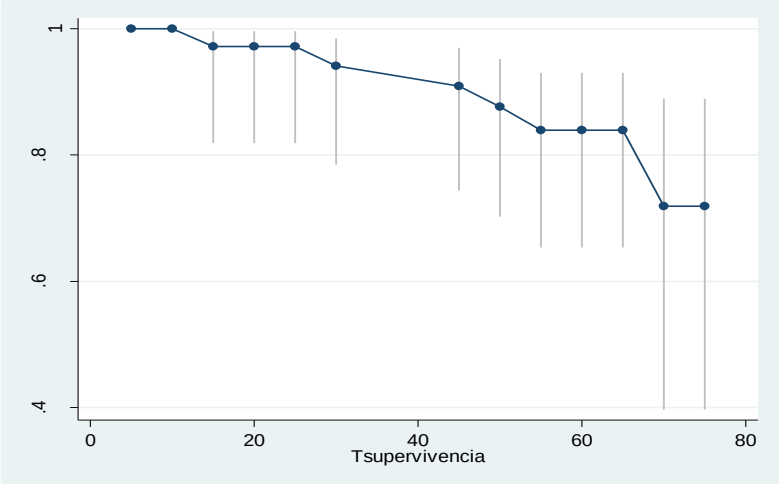


Gráfica 5.1: Función de supervivencia

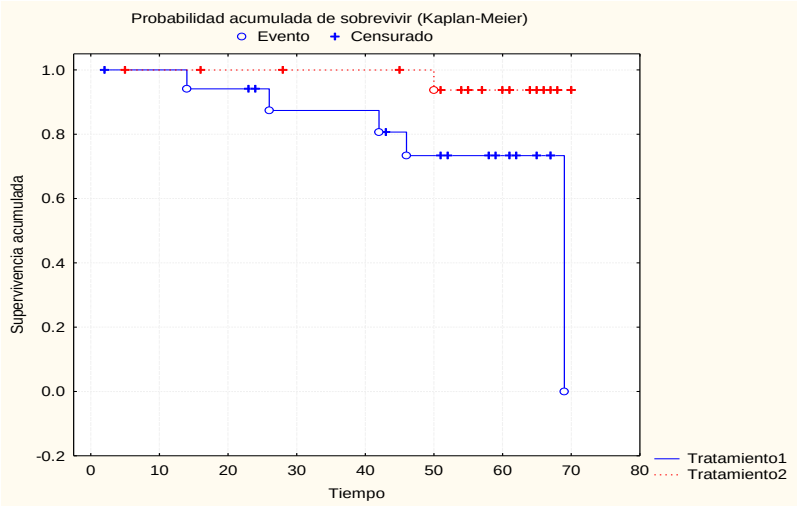
es decir, no importa cual sea el tratamiento que el paciente reciba su probabilidad de sobrevivir es la misma. Para esto se consideró la prueba de log-rank para 2 muestras de supervivencia, la cual se realizó de dos formas; la primera, con los pasos mostrados en el capítulo 3 y la segunda con el resultado del paquete estadístico STATA incluyendo las variables de edad del paciente, tamaño del tumor, nivel de suero de hemoglobina y el índice de Gleason.

La tabla (5.4) muestra el procedimiento de cómo se realiza la prueba de log-rank como se describió en el capítulo 3, la tabla (5.4) da como resultado una $\theta = 3.357833327$ con varianza de 1.475950806 teniendo una $\Omega = 7.639173749$. Dado que el valor de Ω es muy grande y sigue una distribución Ji-cuadrada con un grado de libertad se rechaza la hipótesis nula.

La tabla (5.5) es el resultado obtenido por el paquete estadístico en el cual se incluyen otras variables, ya mencionadas, para el análisis de la prueba log-rank. Además se muestra también la gráfica (5.3) que describe el comportamiento de las funciones de supervivencia para cada tratamiento, estimadas mediante Kaplan-Meier, dado que el resultado de la salida del paquete da un nivel de significancia muy pequeño y por el hecho de que la gráfica de las funciones de supervivencia para cada tratamiento muestra que éstas son en su mayor parte distintas se rechaza la hipótesis nula y se puede concluir que la probabilidad de que un paciente sobreviva más depende fuertemente del tratamiento que reciba.



Gráfica 5.2: Función de supervivencia de la tabla de vida en intervalos de 5 meses



Gráfica 5.3: Función de supervivencia para cada tratamiento

Número de Pacientes	Tratamiento	Tiempo de supervivencia	Censurados	Edad	Nivel del suero de Hemoglobina	Tamaño del tumor	Índice de Gleason
1	1	65	0	67	13.4	34	8
2	2	61	0	60	14.6	4	10
3	2	60	0	77	15.6	3	8
4	1	58	0	64	16.2	6	9
5	2	51	0	65	14.1	21	9
6	1	51	0	61	13.5	8	8
7	1	14	1	73	12.4	18	11
8	1	43	0	60	13.6	7	9
9	2	16	0	73	13.8	8	9
10	1	52	0	73	11.7	5	9
11	1	59	0	77	12.0	7	10
12	2	55	0	74	14.3	7	10
13	2	68	0	71	14.5	19	9
14	2	51	0	65	14.4	10	9
15	1	2	0	76	10.7	8	9
16	1	67	0	70	14.7	7	9
17	2	66	0	70	16.0	8	9
18	2	66	0	70	14.5	15	11
19	2	28	0	75	13.7	19	10
20	2	50	1	68	12.0	20	11
21	1	69	1	60	16.1	26	9
22	1	67	0	71	15.6	8	8
23	2	65	0	51	11.8	2	6
24	1	24	0	71	13.7	10	9
25	2	45	0	72	11.0	4	8
26	2	64	0	74	14.2	4	6
27	1	61	0	75	13.7	10	12
28	1	26	1	72	15.3	37	11
29	1	42	1	57	13.9	24	12
30	2	57	0	72	14.6	8	10
31	2	70	0	72	13.8	3	9
32	2	5	0	74	15.1	3	9
33	2	54	0	51	15.8	7	8
34	1	46	1	72	16.4	4	9
35	2	70	0	73	13.6	2	10
36	2	67	0	68	13.8	7	8
37	1	23	0	63	12.5	2	8
38	1	62	0	69	13.2	3	8

Tabla 5.1: Pacientes con cáncer de próstata

failure _d: Censurados == 1

analysis time _t: Tsupervivencia

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf. Int.]	
2	38	0	1	1.0000	.	.	.
5	37	0	1	1.0000	.	.	.
14	36	1	0	0.9722	0.0274	0.8187	0.9960
16	35	0	1	0.9722	0.0274	0.8187	0.9960
23	34	0	1	0.9722	0.0274	0.8187	0.9960
24	33	0	1	0.9722	0.0274	0.8187	0.9960
26	32	1	0	0.9418	0.0400	0.7865	0.9852
28	31	0	1	0.9418	0.0400	0.7865	0.9852
42	30	1	0	0.9104	0.0495	0.7469	0.9703
43	29	0	1	0.9104	0.0495	0.7469	0.9703
45	28	0	1	0.9104	0.0495	0.7469	0.9703
46	27	1	0	0.8767	0.0580	0.7030	0.9521
50	26	1	0	0.8430	0.0648	0.6618	0.9318
51	25	0	3	0.8430	0.0648	0.6618	0.9318
52	22	0	1	0.8430	0.0648	0.6618	0.9318
54	21	0	1	0.8430	0.0648	0.6618	0.9318
55	20	0	1	0.8430	0.0648	0.6618	0.9318
57	19	0	1	0.8430	0.0648	0.6618	0.9318
58	18	0	1	0.8430	0.0648	0.6618	0.9318
59	17	0	1	0.8430	0.0648	0.6618	0.9318
60	16	0	1	0.8430	0.0648	0.6618	0.9318
61	15	0	2	0.8430	0.0648	0.6618	0.9318
62	13	0	1	0.8430	0.0648	0.6618	0.9318
64	12	0	1	0.8430	0.0648	0.6618	0.9318
65	11	0	2	0.8430	0.0648	0.6618	0.9318
66	9	0	2	0.8430	0.0648	0.6618	0.9318
67	7	0	3	0.8430	0.0648	0.6618	0.9318
68	4	0	1	0.8430	0.0648	0.6618	0.9318
69	3	1	0	0.5620	0.2335	0.0937	0.8691
70	2	0	2	0.5620	0.2335	0.0937	0.8691

Tabla 5.2: Estimación de la función de supervivencia(Kaplan-Meier)

Interval		Beg. Total	Deaths	Lost	Survival	Std. Error	[95% Conf.	Int.]
0	5	38	0	1	1.0000	0.0000	.	.
5	10	37	0	1	1.0000	0.0000	.	.
10	15	36	1	0	0.9722	0.0274	0.8187	0.9960
15	20	35	0	1	0.9722	0.0274	0.8187	0.9960
20	25	34	0	2	0.9722	0.0274	0.8187	0.9960
25	30	32	1	1	0.9414	0.0403	0.7847	0.9850
40	45	30	1	1	0.9094	0.0500	0.7442	0.9700
45	50	28	1	1	0.8764	0.0581	0.7024	0.9519
50	55	26	1	5	0.8391	0.0665	0.6535	0.9302
55	60	20	0	4	0.8391	0.0665	0.6535	0.9302
60	65	16	0	5	0.8391	0.0665	0.6535	0.9302
65	70	11	1	8	0.7192	0.1248	0.3966	0.8892
70	75	2	0	2	0.7192	0.1248	0.3966	0.8892

Tabla 5.3: Tabla de vida para pacientes con cáncer de próstata

Tiempo de supervivencia	d1i	n1i	d2i	n2i	di	ni	E(d1i)=	VAR(1i)=
2	0	18	0	20	0	38	0.0000	0.0000
5	0	18	0	20	0	38	0.0000	0.0000
14	1	18	0	20	1	38	0.4737	0.2493
16	0	17	0	20	0	37	0.0000	0.0000
23	0	17	0	20	0	37	0.0000	0.0000
24	0	17	0	20	0	37	0.0000	0.0000
26	1	17	0	20	1	37	0.4595	0.2484
28	0	16	0	20	0	36	0.0000	0.0000
42	1	16	0	20	1	36	0.4444	0.2469
43	0	15	0	20	0	35	0.0000	0.0000
45	0	15	0	20	0	35	0.0000	0.0000
46	1	15	0	20	1	35	0.4286	0.2449
50	0	14	1	20	1	34	0.4118	0.2422
51	0	14	0	19	0	34	0.0000	0.0000
52	0	14	0	19	0	33	0.0000	0.0000
54	0	14	0	19	0	33	0.0000	0.0000
55	0	14	0	19	0	33	0.0000	0.0000
57	0	14	0	19	0	33	0.0000	0.0000
58	0	14	0	19	0	33	0.0000	0.0000
59	0	14	0	19	0	33	0.0000	0.0000
60	0	14	0	19	0	33	0.0000	0.0000
61	0	14	0	19	0	33	0.0000	0.0000
62	0	14	0	19	0	33	0.0000	0.0000
64	0	14	0	19	0	33	0.0000	0.0000
65	0	14	0	19	0	33	0.0000	0.0000
66	0	14	0	19	0	33	0.0000	0.0000
67	0	14	0	19	0	33	0.0000	0.0000
68	0	14	0	19	0	33	0.0000	0.0000
69	1	14	0	19	1	33	0.4242	0.2443
70	0	13	0	19	0	32	0.0000	0.0000
Total	5		1		6		2.6422	1.4760

Tabla 5.4: Prueba de log-Rank, Primera forma

test Tratamiento, logrank

failure _d: Censurados == 1
analysis time _t: Tsupervivencia

Log-rank test for equality of survivor functions

Tratamiento	Edad	NSueroH	Ttumor	IGleason	Events Observed	Events Expected
1	57	13.9	24	12	1	0.09
1	60	13.6	7	9	0	0.09
1	60	16.1	26	9	1	0.50
1	61	13.5	8	8	0	0.17
1	63	12.5	2	8	0	0.03
1	64	16.2	6	9	0	0.17
1	67	13.4	34	8	0	0.17
1	69	13.2	3	8	0	0.17
1	70	14.7	7	9	0	0.17
1	71	15.6	8	8	0	0.03
1	71	13.7	10	9	0	0.17
1	72	15.3	37	11	1	0.06
1	72	16.4	4	9	1	0.13
1	73	12.4	18	11	0	0.17
1	73	11.7	5	9	1	0.03
1	75	13.7	10	12	0	0.17
1	76	10.7	8	9	0	0.00
1	77	12	7	10	0	0.17
2	51	11.8	2	6	0	0.17
2	51	15.8	7	8	0	0.17
2	60	14.6	4	10	0	0.17
2	65	14.1	21	9	0	0.17
2	65	14.4	10	9	0	0.17
2	68	12	20	11	1	0.17
2	68	13.8	7	8	0	0.17
2	70	16	8	9	0	0.17
2	70	14.5	15	11	0	0.17
2	71	14.5	19	9	0	0.17
2	72	11	4	8	0	0.09
2	72	14.6	8	10	0	0.50
2	72	13.8	3	9	0	0.17
2	73	13.8	8	9	0	0.50
2	73	13.6	2	10	0	0.03
2	74	14.3	7	10	0	0.17
2	74	14.2	4	6	0	0.17
2	74	15.1	3	9	0	0.00
2	75	13.7	19	10	0	0.06
2	77	15.6	3	8	0	0.17
Total					6	6.00

chi2(37)= 76.38
Pr>chi2= 0.0001

Tabla 5.5: Prueba de log-Rank, PE

Capítulo 6

Ejemplo 2. Muestra de cáncer de ovario

Lo que se busco hacer con la siguiente base de datos adjunta en el Apéndice A, fue ajustar un modelo de riesgo proporcional de Cox adecuado para estos datos, en términos de la supervivencia; ya que se desea con este problema estimar la probabilidad de sobrevivir un paciente. Todos los resultados fueron obtenidos del paquete STATISTICA

Se estimarán los coeficientes de regresión para cada covariante y se analizaran qué coeficientes podian estar en el modelo, concluyendo qué variables proporcionan suficiente evidencia para que el modelo se ajuste mejor, basándose en la prueba de Wald. Dado que la base de datos contiene dos diferentes tiempos de supervivencia con sus respectivas censuras, se obtuvo un modelo de Cox para cada variable de tiempo usando los mismos datos correspondientes a las covariantes, y se vio cual fue el que tuvo una mejor significancia para estimar la supervivencia

Primero se realizó la estimación de los parámetros para la variable del tiempo B con su correspondiente censura, el resultado obtenido se presenta en la tabla (6.1), que al igual que la tabla (6.2) se obtuvieron con el paquete STATISTICA, donde se puede apreciar que la prueba de Wald para el coeficiente correspondiente a la textura tuvo un nivel de significancia menor del %5 por lo cual se puede quitar del modelo, de esto se puede concluir que la información que aporta la variable de textura no es tan relevante dentro del modelo, al igual se pudo quitar la variable de *simetría* la cual tiene también una significancia menor del %5, pero en cierto caso no suele ser tan conveniente realizar esta acción, puesto que lo que se hace es quitar la variable menor que se encuentra por debajo del nivel de significancia y se vuelve a realizar el proceso.

Realizando nuevamente el procedimiento se obtiene el resultado descrito en la tabla (6.2), donde se observar que los contrastes que se usan con la estadística de Wald permiten mantener los coeficientes estimados para el modelo de regresión de Cox, el cual es descrito por la gráfica (6.1). Es importante obtener también la función de supervivencia base y no la riesgo debido a que se busca que el modelo de Cox quede en términos de la supervivencia.

	Beta	Error Estándar	t-valor	exponente Beta	Wald	p
Radio	0.0382	0.12334	0.30986	1	0.096014	0.756669
Textura	-0.0844	0.03721	-2.26823	1	5.144846	0.023322
Área	5.8072	25.88240	0.22437	333	0.050341	0.822473
Suavidad	15.9404	9.07956	1.75564	8372171	3.082264	0.079160
Compacidad	0.7850	6.31137	0.12439	2	0.015472	0.901011
Concavidad	1.4046	13.51529	0.10393	4	0.010801	0.917228
Punto de Concavidad	-12.2101	7.30759	-1.67087	0	2.791816	0.094757
Simetría	-123.2675	58.03125	-2.12416		4.512046	0.033665
Diámetro Fractal	-0.1128	0.44050	-0.25614	1	0.065609	0.797843

	Media	Desviación estándar	Mínimo	Máximo
Radio	17.62345	3.47082	10.95000	31.3300
Textura	22.69063	4.54693	10.26000	39.2800
Área	0.10199	0.01257	0.07497	0.1447
Suavidad	0.13756	0.04966	0.04605	0.3114
Compacidad	0.15288	0.07168	0.02398	0.4268
Concavidad	0.08535	0.03401	0.02031	0.2012
Punto de Concavidad	0.19013	0.02679	0.13080	0.3040
Simetría	0.06172	0.00731	0.04655	0.0974
Diámetro Fractal	0.62757	0.39467	0.17060	3.3670

Tabla 6.1: Estimación de los coeficientes para el tiempo B

La función de supervivencia base estimada queda establecida en la tabla (6.3),

Por lo que el modelo de Cox que se considera para el tiempo de supervivencia B con censura su respectiva censura es

$$s(x;t) = \hat{s}(t)^{\exp(0.06X_{1i}+16.65X_{2i}+12.23X_{3i}-1.27X_{4i}+3.81X_{5i}-8.45X_{6i}-110.50X_{7i}-0.11X_{8i})}$$

donde las variables son el radio (X_{1i}), área (X_{2i}), suavidad (X_{3i}), compacidad (X_{4i}), concavidad (X_{5i}), punto de concavidad (X_{6i}), simetría (X_{7i}) y diámetro fractal (X_{8i}).

Se considera ahora el tiempo A con su respectiva censura para ajustar un modelo de regresión de Cox bajo el criterio de la estadística de Wald y las mismas covariantes.

Se realiza la estimación de los parámetros para sus respectivas covariantes. La tabla (6.4) muestra el resultado de la estimación de los parámetros, en la tabla (6.4) se puede observar que los parámetros estimados tienen un valor significativo el cual es adecuado bajo la prueba de Wald, además se puede notar que el valor significativo a la variable *simetría* tiene aproximadamente el valor mínimo para ser aceptado dentro del modelo, pero si se considera una nueva estimación sin tomarla en cuenta se obtiene el resultado descrito por tabla (6.5).

El resultado obtenido en la tabla (6.5) satisface la prueba de Wald para la estimación de todos los parámetros, aunque hay que notar que el valor significativo de éstos tiene cambios significativos, por lo general lo que se busca es emplear el modelo más reducido y con una mejor confiabilidad, ahora el valor significativo total correspondiente a la primera estimación es de .002298 mientras que para el segundo proceso que corresponde al caso sin la *simetría* es de .04994 cuyo valor es mayor, como respuesta un mejor modelo es aquél que tiene una mayor significancia bajo el estidistico Wald, en este caso el modelo sin la

	Beta	Error Estándar	t-valor	exponente Beta	Wald	p
Radio	0.0619	0.12168	0.50862	1	0.258691	0.611025
Área	16.6554	24.95632	0.66738	17114380	0.445400	0.504532
Suavidad	12.2358	8.69672	1.40695	206041	1.979503	0.159452
Compacidad	-1.2767	5.95280	-0.21446	0	0.045994	0.830187
Concavidad	3.8130	13.26563	0.28743	45	0.082618	0.773782
Punto de Concavidad	-8.4578	6.98877	-1.21020	0	1.464574	0.226213
Simetría	-110.5052	56.65569	-1.95047		3.804332	0.051129
Diámetro Fractal	-0.1144	0.44906	-0.25465	1	0.064847	0.798995

	Media	desviación estándar	mínimo	máximo
Radio	17.62345	3.47082	10.95000	31.3300
Área	0.10199	0.01257	0.07497	0.1447
Suavidad	0.13756	0.04966	0.04605	0.3114
Compacidad	0.15288	0.07168	0.02398	0.4268
Concavidad	0.08535	0.03401	0.02031	0.2012
Punto de Concavidad	0.19013	0.02679	0.13080	0.3040
Simetría	0.06172	0.00731	0.04655	0.0974
Diámetro Fractal	0.62757	0.39467	0.17060	3.3670

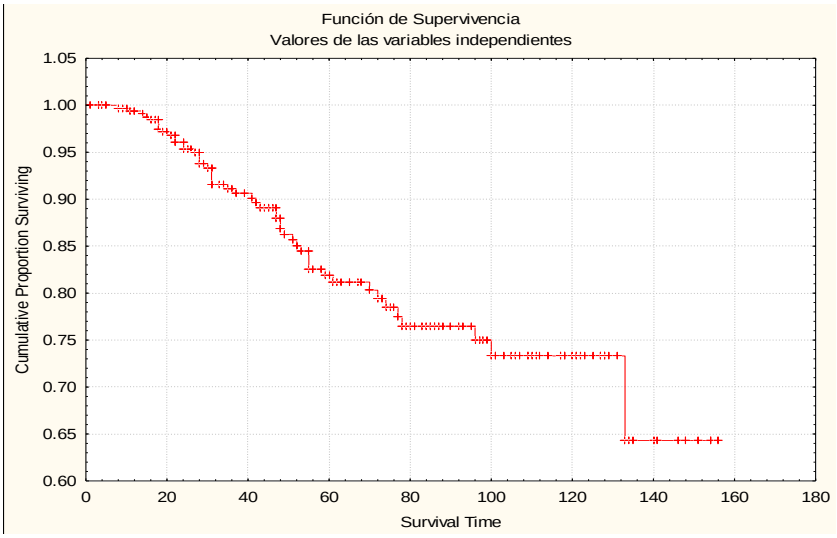
Tabla 6.2: Estimación de los coeficientes sin la variable textura, para el tiempo B

variable résúlto mejor, dado que se utilizaran los valores correspondientes a la estimación sin la *simetría* se calcula la función de supervivencia base la cual que expresada en la tabla (6.6).

Por lo que el modelo de regresión de Cox para el tiempo de supervivencia A con censura correspondiente está dado por

$$s(x;t) = \hat{s}(t)^{\exp(0.07X_{1i} - .03X_{2i} - 8.32X_{3i} - 1.34X_{4i} + 0.49X_{5i} - 8.38X_{6i} - 9.09X_{7i} - 0.29X_{8i})}$$

con el radio (X_{1i}), textura (X_{2i}), área (X_{3i}), suavidad (X_{4i}), compacidad (X_{5i}), concavidad (X_{6i}), punto de concavidad (X_{7i}) y diámetro fractal (X_{8i}).



Gráfica 6.1: Función de supervivencia del modelo de Cox, para el tiempo A

Pacientes	Fun. Sup. Base	Pacientes	Fun. Sup. Base	Pacientes	Fun. Sup. Base
1	0.00003690	46	0.00000000	91	0.00000005
2	0.03105545	47	0.43344082	92	0.00000005
3	0.00012429	48	0.00000000	93	0.00892533
4	0.00001467	49	0.00000233	94	0.00000005
5	0.00003690	50	0.25303876	95	0.00001467
6	0.00383447	51	0.00000000	96	0.00000056
7	0.00682862	52	0.00000821	97	0.00000005
8	0.21089034	53	0.25303876	98	0.00000005
9	0.21089034	54	0.00000056	99	0.00000020
10	0.07699404	55	0.00892533	100	0.00000233
11	1.00000000	56	0.06185462	101	0.00000005
12	0.85268265	57	0.00000000	102	0.00018182
13	0.00003690	58	0.00000005	103	0.00000449
14	0.00001467	59	0.00037842	104	1.00000000
15	0.00000000	60	0.00000000	105	0.00000005
16	0.00000000	61	0.00054370	106	0.00892533
17	0.03105545	62	0.00000115	107	0.00682862
18	0.00000005	63	0.00211245	108	0.00000005
19	0.00000020	64	0.00000005	109	0.00000449
20	0.00000115	65	0.00000005	110	0.00000020
21	0.00000000	66	0.00000005	111	0.00054370
22	0.00000056	67	0.72273944	112	0.00000005
23	0.00000000	68	0.00000005	113	0.00001467
24	0.00000000	69	0.00001467	114	0.00000005
25	0.00000005	70	0.00000005	115	0.00000005
26	0.00000005	71	0.00211245	116	0.00001467
27	0.00000000	72	0.00000000	117	0.00000005
28	0.11658185	73	0.00002344	118	0.00000056
29	0.00012429	74	0.00000005	119	0.00000821
30	0.00000020	75	0.21089034	120	0.00000005
31	0.00000000	76	0.00000056	121	0.00000020
32	0.00000005	77	0.00000056	122	0.00003690
33	0.43344082	78	0.00000005	123	0.21089034
34	0.00018182	79	0.00211245	124	0.00000005
35	0.00000005	80	0.00054370	125	0.00000005
36	0.00000000	81	0.85268265	126	0.00000056
37	0.00000056	82	0.00682862	127	0.00000005
38	0.07699404	83	0.00000005	128	0.00000056
39	0.00001467	84	0.00000056	129	0.00003690
40	0.00000005	85	0.00108267	130	0.00026276
41	0.72273944	86	0.00000005	131	0.02435592
42	0.00892533	87	0.03105545	132	0.00018182
43	0.00000056	88	0.00001467	133	0.00000020
44	0.00000000	89	1.00000000	134	0.07699404
45	0.00892533	90	0.00000005	135	0.00000056

Pacientes	Fun. Sup. Base	Pacientes	Fun. Sup. Base	Pacientes	Fun. Sup. Base
136	0.25303876	181	0.00000449	226	0.00892533
137	0.85268265	182	0.11658185	227	0.00514322
138	0.00001467	183	0.00211245	228	0.00514322
139	1.00000000	184	0.51474797	229	0.25303876
140	0.00000449	185	0.00002344	230	0.06185462
141	0.00000233	186	0.00001467	231	1.00000000
142	0.00000005	187	0.00001467	232	0.43344082
143	0.00000056	188	0.00001467	233	0.21089034
144	0.00000020	189	0.00000449	234	0.00892533
145	0.00000056	190	0.00000821	235	0.07699404
146	0.00000005	191	1.00000000	236	1.00000000
147	0.00892533	192	0.00211245	237	0.03105545
148	0.00892533	193	0.00514322	238	0.06185462
149	0.00514322	194	0.43344082	239	0.11658185
150	0.00003690	195	0.00001467	240	0.06185462
151	0.00000005	196	0.00012429	241	0.03105545
152	0.03105545	197	0.43344082	242	0.11658185
153	0.00001467	198	0.00003690	243	0.43344082
154	0.00000056	199	0.00003690	244	0.85268265
155	0.00000056	200	0.00001467	245	1.00000000
156	0.00003690	201	0.07699404	246	0.85268265
157	0.00211245	202	0.00003690	247	0.43344082
158	0.00000056	203	0.00054370	248	1.00000000
159	0.00001467	204	0.00211245	249	0.17339804
160	0.00000056	205	0.00003690	250	0.21089034
161	0.00285226	206	0.00514322	251	1.00000000
162	0.00000056	207	0.00054370	252	1.00000000
163	0.00001467	208	0.00018182		
164	0.00000056	209	0.00211245		
165	0.00000056	210	0.07699404		
166	0.00012429	211	0.00383447		
167	0.00000056	212	0.00211245		
168	0.85268265	213	0.00211245		
169	0.00000449	214	0.00026276		
170	0.00108267	215	0.07699404		
171	0.00108267	216	0.85268265		
172	0.00000449	217	0.07699404		
173	0.00002344	218	0.07699404		
174	0.43344082	219	0.00514322		
175	0.00000233	220	0.03105545		
176	0.72273944	221	0.00892533		
177	0.00108267	222	0.07699404		
178	0.17339804	223	0.07699404		
179	0.00002344	224	0.61045459		
180	0.00682862	225	0.03105545		

Tabla 6.3: Función de supervivencia base para el tiempo B

	Beta	Error Estándar	t-valor	Exponente	Wald	p
Radio	0.0003	0.10307	0.00278	1.000	0.000008	0.997785
Textura	-0.0380	0.02993	-1.26934	0.963	1.611214	0.204331
Área	4.7516	20.87866	0.22758	115.774	0.051794	0.819971
Suavidad	8.5588	7.44509	1.14959	5212.395	1.321557	0.250321
Compacidad	-1.2502	5.06922	-0.24663	0.286	0.060829	0.805193
Concavidad	8.9122	11.89249	0.74939	7421.620	0.561591	0.453626
Punto de Concavidad	-9.4525	6.28515	-1.50394	0.000	2.261830	0.132607
Simetría	-90.6382	46.83740	-1.93517		3.744870	0.052979
Diámetro Fractal	-0.1895	0.39839	-0.47575	0.827	0.226334	0.634259

	Media	Dev. Estándar	mínimo	máximo
Radio	17.62345	3.47082	10.95000	31.3300
Textura	22.69063	4.54693	10.26000	39.2800
Área	0.10199	0.01257	0.07497	0.1447
Suavidad	0.13756	0.04966	0.04605	0.3114
Compacidad	0.15288	0.07168	0.02398	0.4268
Concavidad	0.08535	0.03401	0.02031	0.2012
Punto de Concavidad	0.19013	0.02679	0.13080	0.3040
Simetría	0.06172	0.00731	0.04655	0.0974
Diámetro Fractal	0.62757	0.39467	0.17060	3.3670

Tabla 6.4: Estimación de los parámetros, para el tiempo B

	Beta	Error Estándar	t-valor	exponente	Wald	p
Radio	0.07419	0.09688	0.76583	1.077	0.586500	0.443782
Textura	-0.03242	0.02931	-1.10595	0.968	1.223126	0.268756
Área	-8.32272	20.11146	-0.41383	0.000	0.171255	0.679002
Suavidad	-1.34758	5.56027	-0.24236	0.260	0.058737	0.808505
Compacidad	0.49132	5.12259	0.09591	1.634	0.009199	0.923590
Concavidad	8.38398	12.10947	0.69235	4376.399	0.479348	0.488723
Punto de Concavidad	-9.09729	6.23318	-1.45949	0.000	2.130123	0.144439
Diámetro Fractal	-0.29546	0.40087	-0.73703	0.744	0.543212	0.461110

	Media	Dev. Estándar	mínimo	máximo
Radio	17.62345	3.47082	10.95000	31.3300
Textura	22.69063	4.54693	10.26000	39.2800
Área	0.10199	0.01257	0.07497	0.1447
Suavidad	0.13756	0.04966	0.04605	0.3114
Compacidad	0.15288	0.07168	0.02398	0.4268
Concavidad	0.08535	0.03401	0.02031	0.2012
Punto de Concavidad	0.19013	0.02679	0.13080	0.3040
Diámetro Fractal	0.62757	0.39467	0.17060	3.3670

Tabla 6.5: Estimación de los parámetros sin considerar él correspondiente a la variable simetría, para el tiempo A

Pacientes	Fun. Sup. Base	Pacientes	Fun. Sup. Base	Pacientes	Fun. Sup. Base
1	0.23008647	46	0.08757061	91	0.08757061
2	0.37924677	47	0.76913807	92	0.11248594
3	0.24029093	48	0.08757061	93	0.57029498
4	0.63693102	49	0.17389088	94	0.11248594
5	0.23008647	50	0.80049507	95	0.2097007
6	0.39037908	51	0.08757061	96	0.14157219
7	0.33438412	52	0.26081915	97	0.08757061
8	0.47082521	53	0.67909997	98	0.08757061
9	0.45891102	54	0.14157219	99	0.17389088
10	0.42435748	55	0.3565416	100	0.17389088
11	0.88227586	56	0.47082521	101	0.11248594
12	0.72337959	57	0.08757061	102	0.25053092
13	0.23008647	58	0.08757061	103	0.18574205
14	0.2097007	59	0.39037908	104	0.88227586
15	0.08757061	60	0.08757061	105	0.11248594
16	0.08757061	61	0.55711321	106	0.57029498
17	0.37924677	62	0.30225432	107	0.34541556
18	0.15171111	63	0.28152722	108	0.11248594
19	0.12725337	64	0.11248594	109	0.18574205
20	0.15171111	65	0.11248594	110	0.12725337
21	0.08757061	66	0.08757061	111	0.26081915
22	0.14157219	67	0.72337959	112	0.11248594
23	0.08757061	68	0.08757061	113	0.2097007
24	0.08757061	69	0.2097007	114	0.11248594
25	0.11248594	70	0.08757061	115	0.11248594
26	0.33438412	71	0.30225432	116	0.2097007
27	0.08757061	72	0.08757061	117	0.11248594
28	0.72337959	73	0.34541556	118	0.14157219
29	0.24029093	74	0.08757061	119	0.19782182
30	0.14157219	75	0.47082521	120	0.11248594
31	0.08757061	76	0.18574205	121	0.12725337
32	0.08757061	77	0.14157219	122	0.21995182
33	0.59663508	78	0.08757061	123	0.47082521
34	0.25053092	79	0.28152722	124	0.12725337
35	0.11248594	80	0.84887532	125	0.11248594
36	0.08757061	81	0.72337959	126	0.14157219
37	0.14157219	82	0.33438412	127	0.11248594
38	0.42435748	83	0.08757061	128	0.14157219
39	0.2097007	84	0.14157219	129	0.31279982
40	0.11248594	85	0.29177321	130	0.43582847
41	0.63693102	86	0.11248594	131	0.63693102
42	0.37924677	87	0.80049507	132	0.3235454
43	0.14157219	88	0.2097007	133	0.14157219
44	0.08757061	89	0.88227586	134	0.88227586
45	0.3565416	90	0.11248594	135	0.14157219

Pacientes	Fun. Sup. Base	Pacientes	Fun. Sup. Base	Pacientes	Fun. Sup. Base
136	0.67909997	181	0.18574205	226	0.47082521
137	0.80049507	182	0.67909997	227	0.3235454
138	0.2097007	183	0.28152722	228	0.3235454
139	0.96634199	184	0.84887532	229	0.55711321
140	0.21995182	185	0.21995182	230	0.37924677
141	0.3565416	186	0.2097007	231	0.88227586
142	0.11248594	187	0.2097007	232	0.59663508
143	0.14157219	188	0.2097007	233	0.47082521
144	0.14157219	189	0.19782182	234	0.47082521
145	0.14157219	190	0.19782182	235	0.39037908
146	0.14157219	191	0.93269189	236	0.96634199
147	0.3565416	192	0.29177321	237	0.37924677
148	0.44724882	193	0.31279982	238	0.37924677
149	0.42435748	194	0.57029498	239	0.43582847
150	0.36775524	195	0.2097007	240	0.37924677
151	0.12725337	196	0.24029093	241	0.37924677
152	0.91582114	197	0.59663508	242	0.43582847
153	0.88227586	198	0.21995182	243	0.57029498
154	0.14157219	199	0.21995182	244	0.76913807
155	0.14157219	200	0.2097007	245	0.93269189
156	0.24029093	201	0.60986201	246	0.76913807
157	0.93269189	202	0.96634199	247	0.57029498
158	0.14157219	203	0.26081915	248	0.91582114
159	0.2097007	204	0.28152722	249	0.80049507
160	0.14157219	205	0.23008647	250	0.47082521
161	0.47082521	206	0.3235454	251	0.93269189
162	0.14157219	207	0.26081915	252	0.96634199
163	0.23008647	208	0.25053092		
164	0.19782182	209	0.30225432		
165	0.14157219	210	0.42435748		
166	0.25053092	211	0.30225432		
167	0.14157219	212	0.28152722		
168	0.96634199	213	0.28152722		
169	0.19782182	214	0.47082521		
170	0.76913807	215	0.39037908		
171	0.27114942	216	0.80049507		
172	0.2097007	217	0.39037908		
173	0.21995182	218	0.39037908		
174	0.59663508	219	0.39037908		
175	0.27114942	220	0.37924677		
176	0.93269189	221	0.37924677		
177	0.27114942	222	0.42435748		
178	0.60986201	223	0.42435748		
179	0.21995182	224	0.72337959		
180	0.63693102	225	0.37924677		

Tabla 6.6: Función de supervivencia base para el tiempo A

Capítulo 7

Conclusión

Se ha visto en este trabajo quizá no con tanta profundidad lo que es la teoría general de lo que es el análisis de supervivencia, pero si con un grado matemático y con un enfoque teórico-práctico que asume una familiaridad con la probabilidad y estadística. Durante el desarrollo del mismo se han mostrado resultados de suma importancia para la comprensión del manejo de los datos que se usan para la elaboración de un análisis de supervivencia como lo son el estimador Kaplan-Meier o el modelo de Cox por citar algunos. Se rescata la importancia del análisis de supervivencia en áreas médicas, sociales e industriales entre otras, la cual representa un claro ejemplo de la utilidad de los procedimientos estadísticos para resolver problemas de investigación.

Podría surgir de manera natural después de haber leído el presente trabajo una respuesta más clara a la interrogante ¿para qué sirve el análisis de supervivencia? dicha solución se podría argumentar diciendo que sirve para resolver problemas del tipo que se presentan en la medicina entre otras, debido a que es, y así se ha mostrado en el desarrollo del mismo, un conjunto de herramientas o técnicas estadísticas que se emplean para analizar ciertos tipos de datos conocidos como de "cohorte", además de que estás técnicas no son las habituales que se usan en los métodos estadísticos para variables cuantitativas, pero si buscan tener una similaridad a los métodos habituales, como análisis descriptivo o modelos de tipo regresión, etcétera, para conseguir un comprensión más sencilla de lo que representan los datos.

Existen diversos problemas que surgen durante el desarrollo de un análisis en supervivencia, como de los más usuales es poder registrar correctamente los datos censurados y definir bien lo que en el problema se reconoce como censura, que los datos registrados sean difíciles de manejar, y quizá requieran de una transformación adecuada para poder aplicar los métodos mostrados en está tesis.

Se desea que el desarrollo de este trabajo proporcione al estudiante, o aquella persona interesada en el tema, herramientas que le permitan buscar u obtener una solución verosímil de los problemas donde la investigación estadística maneje datos censurados o cohorte y también tendrá una visión más amplia de lo que se puede aplicar en campo de

la estadística.

Apéndice A

Base de datos para muestra de cáncer de ovario

La siguiente tabla muestra la base de datos que corresponde al ejemplo dos, muestra de cáncer de ovario. Donde el tiempo A esta medido en meses y representa la supervivencia libre de enfermedades y el lapso de tiempo en recaen en la enfermedad, al igual que el tiempo B esta medido de la misma forma y representa a los pacientes que mueren, viven o sobreviven los 158 meses que dura el estudio. La censura A representa a los paciente que no recaen y se les asigna el valor cero, mientra que los que recaen se les asigna el valor uno. La censura B representa a los pacientes que mueren por causas ajenas al cancer o aquellas del grupo de estudio que sobreviven y se les asigna el valor cero, mientras que las que mueren a causa del cancer se les asigna el valor 1.

Pacientes	CensuraA	CensuraB	TiempoA	TiempoB	Radio	Textura	Aréa
1	0	0	55	55	18.02000	27.60000	0.09489
2	0	0	29	29	14.09000	27.70000	0.11860
3	0	0	53	53	13.32000	27.25000	0.08446
4	1	0	12	68	15.11000	32.79000	0.09656
5	0	0	56	56	13.99000	36.52000	0.11780
6	1	1	26	41	15.54000	31.46000	0.09863
7	0	0	36	36	15.77000	27.48000	0.10910
8	1	0	19	19	19.00000	23.31000	0.10060
9	1	0	20	20	16.09000	28.58000	0.09909
10	0	0	24	24	14.41000	26.57000	0.09666
11	0	0	5	5	17.43000	17.67000	0.07715
12	0	0	10	10	26.50000	27.47000	0.08681
13	0	0	56	56	16.25000	25.27000	0.10840
14	0	0	61	61	17.99000	10.38000	0.11840
15	0	0	154	154	21.37000	17.44000	0.08836
16	0	0	148	148	11.42000	20.38000	0.14250
17	1	1	27	28	20.29000	14.34000	0.10030
18	1	1	77	100	12.75000	15.29000	0.11890
19	0	0	99	99	18.98000	19.61000	0.09087
20	1	0	77	77	13.71000	20.83000	0.11890
21	0	0	156	156	13.00000	21.82000	0.12730
22	0	0	79	79	12.46000	24.04000	0.11860
23	0	0	154	154	16.02000	23.24000	0.08206
24	0	0	156	156	15.78000	17.89000	0.09710
25	0	0	117	117	15.85000	23.95000	0.08401
26	1	0	36	114	14.54000	27.54000	0.11390
27	0	0	154	154	14.68000	20.13000	0.09867
28	1	1	10	22	16.13000	20.68000	0.11700
29	0	0	53	53	15.34000	14.26000	0.10730
30	0	0	96	96	16.65000	21.38000	0.11210
31	0	0	146	146	17.14000	16.40000	0.11860
32	0	0	128	128	14.58000	21.53000	0.10540
33	0	0	16	16	18.61000	20.25000	0.09440
34	0	0	52	52	15.30000	25.27000	0.10820
35	0	0	118	118	17.57000	15.05000	0.09847
36	0	0	140	140	11.84000	18.70000	0.11090
37	0	0	87	87	17.02000	23.98000	0.11970
38	0	0	25	25	16.53000	20.71000	0.10800
39	0	0	65	65	19.27000	26.47000	0.09401
40	0	0	109	109	16.13000	17.88000	0.10400
41	0	0	12	12	16.74000	21.59000	0.09610
42	0	0	31	31	14.25000	21.72000	0.09823
43	0	0	80	80	14.99000	25.20000	0.09387

Pacientes	CensuraA	CensuraB	TiempoA	TiempoB	Radio	Textura	Aréa
44	1	1	121	133	13.48000	20.82000	0.10160
45	0	0	34	34	13.44000	21.58000	0.08162
46	0	0	135	135	10.95000	21.35000	0.12270
47	1	1	9	16	19.07000	24.81000	0.09081
48	0	0	133	133	13.28000	20.28000	0.10410
49	0	0	76	76	13.17000	21.81000	0.09714
50	1	1	8	18	18.65000	17.60000	0.10990
51	0	0	151	151	13.17000	18.66000	0.11580
52	1	1	48	70	15.10000	22.02000	0.09056
53	1	1	11	18	19.21000	18.57000	0.10530
54	0	0	85	85	14.71000	21.59000	0.11370
55	1	0	34	34	14.25000	22.15000	0.10490
56	1	1	19	27	12.68000	23.84000	0.11220
57	0	0	141	141	14.78000	23.94000	0.11720
58	0	0	123	123	17.20000	24.52000	0.10710
59	1	1	26	49	13.80000	15.79000	0.10070
60	0	0	133	133	19.55000	15.49000	0.10790
61	1	1	18	48	20.04000	15.52000	0.08543
62	1	1	40	77	19.10000	26.29000	0.12150
63	0	0	45	45	18.08000	19.74000	0.10620
64	0	0	118	118	14.48000	21.46000	0.09444
65	0	0	118	118	19.02000	24.59000	0.09029
66	0	0	128	128	15.06000	19.83000	0.10390
67	1	1	10	11	20.77000	22.83000	0.10330
68	0	0	129	129	14.42000	19.77000	0.09752
69	0	0	68	68	14.19000	26.02000	0.10500
70	0	0	125	125	13.11000	15.56000	0.13980
71	0	0	43	43	14.87000	16.67000	0.11620
72	0	0	134	134	15.78000	22.91000	0.11550
73	1	1	35	59	17.95000	20.01000	0.08402
74	0	0	131	131	18.66000	17.12000	0.10540
75	0	0	19	19	24.25000	20.20000	0.14470
76	1	1	73	78	19.00000	18.91000	0.08217
77	0	0	88	88	19.79000	25.12000	0.10150
78	0	0	128	128	16.16000	21.54000	0.10080
79	1	0	45	45	15.71000	13.93000	0.09462
80	1	1	7	48	20.29000	21.49000	0.09258
81	0	0	10	10	12.77000	22.47000	0.09055
82	0	0	36	36	14.95000	17.57000	0.11670
83	0	0	128	128	16.11000	18.05000	0.09721
84	0	0	85	85	11.80000	16.58000	0.10910
85	1	1	44	47	17.68000	20.74000	0.11150
86	1	0	101	121	19.89000	16.89000	0.10050
87	1	1	8	28	19.59000	18.15000	0.11200

Pacientes	CensuraA	CensuraB	TiempoA	TiempoB	Radio	Textura	Area
88	0	0	62	62	23.27000	22.04000	0.08439
89	0	0	5	5	16.78000	18.80000	0.08865
90	0	0	109	109	18.98000	24.12000	0.11980
91	0	0	127	127	13.43000	19.63000	0.09048
92	0	0	120	120	15.46000	11.89000	0.12570
93	1	1	17	31	27.22000	21.87000	0.10940
94	0	0	109	109	21.09000	26.57000	0.11410
95	0	0	65	65	15.70000	20.31000	0.09597
96	0	0	81	81	15.28000	22.41000	0.09057
97	0	0	127	127	18.31000	18.58000	0.08588
98	0	0	122	122	14.22000	23.12000	0.10750
99	1	1	74	96	12.34000	26.86000	0.10340
100	0	0	75	75	14.86000	23.21000	0.10440
101	0	0	117	117	13.77000	22.29000	0.12000
102	0	0	52	52	19.18000	22.49000	0.08523
103	0	0	73	73	14.45000	20.22000	0.09872
104	0	0	5	5	23.29000	26.67000	0.11410
105	0	0	110	110	13.81000	23.75000	0.13230
106	1	1	17	31	15.12000	16.68000	0.08876
107	0	0	35	35	28.11000	18.47000	0.11420
108	0	0	105	105	17.42000	25.56000	0.10060
109	0	0	73	73	14.19000	23.81000	0.09463
110	0	0	98	98	13.86000	16.93000	0.10260
111	0	0	48	48	19.80000	21.56000	0.09383
112	0	0	111	111	19.53000	32.47000	0.08420
113	0	0	68	68	16.34000	20.81000	0.10310
114	0	0	103	103	12.83000	22.33000	0.10880
115	0	0	112	112	17.05000	19.08000	0.11410
116	0	0	67	67	20.51000	27.81000	0.09159
117	0	0	107	107	23.21000	26.97000	0.09509
118	0	0	83	83	17.46000	39.28000	0.09812
119	0	0	70	70	19.40000	23.50000	0.10270
120	0	0	106	106	17.30000	17.08000	0.10080
121	0	0	97	97	19.45000	19.33000	0.10350
122	0	0	58	58	13.96000	17.05000	0.10960
123	1	1	19	19	19.55000	28.77000	0.09260
124	0	0	100	100	16.46000	23.69000	0.10180
125	0	0	101	101	15.32000	17.27000	0.13350
126	0	0	93	93	17.77000	10.26000	0.11060
127	0	0	109	109	15.66000	23.20000	0.11090
128	0	0	81	81	15.53000	33.56000	0.10630
129	1	1	39	55	20.31000	27.06000	0.10000
130	1	1	22	51	17.62000	22.69000	0.09811
131	1	1	12	30	17.29000	22.13000	0.08999

Pacientes	CensuraA	CensuraB	TiempoA	TiempoB	Radio	Textura	Area
132	1	1	37	52	17.19000	22.07000	0.09726
133	0	0	96	96	20.73000	31.12000	0.09469
134	1	1	5	24	21.75000	20.99000	0.09401
135	0	0	80	80	17.93000	24.48000	0.08855
136	1	1	11	18	19.71000	19.06000	0.10180
137	0	0	8	8	16.24000	18.77000	0.10660
138	0	0	68	68	16.31000	25.03000	0.09307
139	0	0	1	1	11.76000	18.14000	0.09968
140	1	1	58	72	19.53000	18.90000	0.11500
141	0	1	34	74	20.09000	23.86000	0.10800
142	0	0	101	101	18.22000	18.87000	0.09746
143	0	0	90	90	12.74000	19.36000	0.10490
144	0	0	96	96	20.16000	19.66000	0.08020
145	0	0	95	95	20.34000	21.51000	0.11700
146	1	0	78	121	16.27000	20.71000	0.11690
147	0	0	34	34	17.06000	21.00000	0.11190
148	1	1	21	31	18.77000	21.43000	0.09116
149	1	1	24	37	23.51000	24.27000	0.10690
150	1	1	33	55	19.68000	21.68000	0.09797
151	1	0	97	100	15.75000	19.22000	0.12430
152	1	1	4	28	25.73000	17.46000	0.11490
153	1	0	5	63	15.08000	25.74000	0.10240
154	0	0	86	86	20.44000	21.78000	0.09150
155	0	0	78	78	20.20000	26.83000	0.09905
156	1	1	53	55	31.33000	21.84000	0.09971
157	1	1	2	43	22.01000	21.90000	0.10630
158	0	0	85	85	20.64000	17.35000	0.09446
159	0	0	63	63	11.08000	18.83000	0.12160
160	0	0	86	86	14.60000	23.29000	0.08682
161	1	1	19	42	19.55000	23.21000	0.10100
162	0	0	83	83	21.61000	22.28000	0.11670
163	1	1	54	61	17.91000	21.02000	0.12300
164	1	0	70	92	17.99000	20.66000	0.10360
165	0	0	84	84	15.13000	29.81000	0.08320
166	1	1	49	53	15.50000	21.08000	0.11200
167	0	0	83	83	17.29000	21.43000	0.10040
168	1	1	1	8	20.18000	19.54000	0.11330
169	0	0	72	72	18.82000	21.97000	0.10180
170	1	1	9	47	13.98000	19.62000	0.10600
171	0	0	47	47	17.27000	25.42000	0.08331
172	1	0	61	73	18.34000	28.31000	0.08067
173	0	0	59	59	17.08000	27.15000	0.09898
174	1	0	16	16	17.75000	28.03000	0.09997
175	1	0	47	75	21.10000	20.52000	0.09684

Pacientes	CensuraA	CensuraB	TiempoA	TiempoB	Radio	Textura	Area
176	1	0	2	12	19.59000	25.00000	0.10320
177	0	0	47	47	17.60000	23.33000	0.09289
178	1	1	14	21	19.44000	18.82000	0.10890
179	0	0	60	60	16.69000	20.20000	0.07497
180	1	1	12	35	18.01000	20.56000	0.10010
181	0	0	73	73	18.49000	17.52000	0.10120
182	1	1	11	22	20.59000	21.24000	0.10850
183	0	0	46	46	13.82000	24.49000	0.11620
184	1	1	7	15	24.24000	18.74000	0.08938
185	0	0	60	60	15.46000	23.95000	0.11830
186	0	0	63	63	15.05000	19.07000	0.09215
187	0	0	68	68	18.31000	20.58000	0.10680
188	0	0	68	68	19.89000	20.26000	0.10370
189	0	0	72	72	24.63000	21.60000	0.10300
190	0	0	70	70	14.27000	22.55000	0.10380
191	0	0	3	3	18.45000	31.39000	0.09889
192	0	0	44	44	15.91000	17.22000	0.13940
193	0	0	39	39	15.22000	30.62000	0.10480
194	0	0	17	17	20.92000	25.09000	0.10990
195	0	0	61	61	21.56000	22.39000	0.11100
196	0	0	53	53	20.13000	28.25000	0.09780
197	0	0	16	16	16.60000	28.08000	0.08455
198	0	0	58	58	13.86000	23.83000	0.10670
199	0	0	58	58	18.11000	26.17000	0.09867
200	0	0	63	63	21.93000	30.64000	0.08679
201	1	1	14	24	17.53000	25.28000	0.09278
202	1	0	1	58	18.11000	30.99000	0.08625
203	0	0	48	48	24.29000	25.48000	0.09374
204	0	0	45	45	16.48000	27.01000	0.09853
205	0	0	55	55	15.60000	26.79000	0.07885
206	0	0	37	37	15.78000	17.10000	0.09668
207	0	0	48	48	19.28000	20.88000	0.09033
208	0	0	52	52	15.66000	24.51000	0.08886
209	0	0	43	43	22.44000	27.42000	0.12110
210	0	0	24	24	17.98000	23.96000	0.11570
211	0	0	41	41	13.63000	24.70000	0.10550
212	0	0	45	45	23.01000	33.87000	0.11570
213	0	0	45	45	22.41000	29.95000	0.11190
214	1	0	19	51	12.53000	30.98000	0.09252
215	0	0	26	26	19.80000	20.46000	0.09652
216	0	0	8	8	19.96000	27.41000	0.09075
217	0	0	26	26	13.84000	23.20000	0.11640
218	0	0	26	26	14.90000	23.50000	0.09757
219	1	0	26	39	17.87000	18.47000	0.08684

Pacientes	CensuraA	CensuraB	TiempoA	TiempoB	Radio	Textura	Area
220	0	0	29	29	19.22000	27.18000	0.10900
221	0	0	31	31	14.72000	25.26000	0.11740
222	0	0	25	25	17.99000	24.44000	0.07868
223	0	0	25	25	17.99000	24.44000	0.07868
224	1	1	10	14	20.35000	23.95000	0.11440
225	0	0	28	28	15.88000	22.52000	0.10560
226	1	1	19	31	22.52000	21.92000	0.07592
227	0	0	37	37	15.44000	31.18000	0.09399
228	0	0	37	37	17.17000	29.19000	0.08952
229	0	0	18	18	16.70000	28.13000	0.08896
230	0	0	27	27	13.18000	27.96000	0.11830
231	0	0	5	5	20.60000	27.65000	0.10260
232	0	0	16	16	15.01000	30.75000	0.09865
233	0	0	19	19	29.86000	22.41000	0.08661
234	1	0	19	33	23.66000	19.26000	0.10040
235	0	0	26	26	18.90000	25.90000	0.08560
236	0	0	1	1	19.80000	21.17000	0.08148
237	0	0	28	28	20.69000	27.84000	0.09430
238	0	0	27	27	24.90000	31.90000	0.10940
239	0	0	22	22	24.53000	29.25000	0.11700
240	0	0	27	27	16.76000	25.83000	0.09311
241	0	0	28	28	15.42000	28.10000	0.10290
242	0	0	22	22	17.83000	24.73000	0.10070
243	0	0	17	17	19.32000	24.44000	0.10100
244	0	0	9	9	21.67000	21.22000	0.09883
245	0	0	3	3	23.48000	28.36000	0.08082
246	0	0	9	9	14.41000	27.93000	0.09715
247	0	0	17	17	19.05000	18.99000	0.09437
248	0	0	4	4	25.62000	23.93000	0.09954
249	1	0	8	21	16.03000	26.45000	0.09198
250	0	0	19	19	21.56000	14.79000	0.08964
251	0	0	3	3	17.54000	23.94000	0.09367
252	0	0	1	1	24.14000	13.70000	0.08801

Pacientes	Suavidad	Compacidad	Concavidad	Punto de Concavidad	Simetria	Diametro Fractal
1	0.10360	0.10860	0.07055	0.18650	0.06333	0.62490
2	0.07879	0.05168	0.02752	0.16580	0.05879	0.20900
4	0.12230	0.13380	0.05728	0.16970	0.06116	0.30740
5	0.14060	0.10720	0.04155	0.15030	0.06268	0.47800
6	0.10350	0.10450	0.05742	0.19290	0.06038	0.31520
7	0.15070	0.19480	0.07333	0.15430	0.06284	0.63930
8	0.11840	0.13590	0.09610	0.20080	0.05618	0.61770
9	0.17990	0.17650	0.07776	0.19560	0.06411	0.31870
10	0.16100	0.13470	0.05563	0.16020	0.06794	0.58980
11	0.06327	0.04027	0.03278	0.14900	0.05151	0.44960
12	0.10070	0.15300	0.11120	0.18410	0.05203	2.08400
13	0.10620	0.11060	0.06908	0.18500	0.05973	0.39360
14	0.27760	0.30010	0.14710	0.24190	0.07871	1.09500
15	0.11890	0.12550	0.08180	0.23330	0.06010	0.58540
16	0.28390	0.24140	0.10520	0.25970	0.09744	0.49560
17	0.13280	0.19800	0.10430	0.18090	0.05883	0.75720
18	0.15690	0.16640	0.07666	0.19950	0.07164	0.38770
19	0.12370	0.12130	0.08910	0.17270	0.05767	0.52850
20	0.16450	0.09366	0.05985	0.21960	0.07451	0.58350
21	0.19320	0.18590	0.09353	0.23500	0.07389	0.30630
22	0.23960	0.22730	0.08543	0.20300	0.08243	0.29760
23	0.06669	0.03299	0.03323	0.15280	0.05697	0.37950
24	0.12920	0.09954	0.06606	0.18420	0.06082	0.50580
25	0.10020	0.09938	0.05364	0.18470	0.05338	0.40330
26	0.15950	0.16390	0.07364	0.23030	0.07077	0.37000
27	0.07200	0.07395	0.05259	0.15860	0.05922	0.47270
28	0.20220	0.17220	0.10280	0.21640	0.07356	0.56920
29	0.21350	0.20770	0.09756	0.25210	0.07032	0.43880
30	0.14570	0.15250	0.09170	0.19950	0.06330	0.80680
31	0.22760	0.22290	0.14010	0.30400	0.07413	1.04600
32	0.10660	0.14900	0.07731	0.16970	0.05699	0.85290
34	0.16970	0.16830	0.08751	0.19260	0.06540	0.43900
35	0.11570	0.09875	0.07953	0.17390	0.06149	0.60030
36	0.15160	0.12180	0.05182	0.23010	0.07799	0.48250
37	0.14960	0.24170	0.12030	0.22480	0.06382	0.60090
38	0.19670	0.13960	0.09196	0.19320	0.06180	0.94410
39	0.17190	0.16570	0.07593	0.18530	0.06261	0.55580
40	0.15590	0.13540	0.07752	0.19980	0.06515	0.33400
41	0.13360	0.13480	0.06018	0.18960	0.05656	0.46150
42	0.10980	0.13190	0.05598	0.18850	0.06125	0.28600
43	0.05131	0.02398	0.02899	0.15650	0.05504	1.21400
44	0.12550	0.10630	0.05439	0.17200	0.06419	0.21300

Pacientes	Suavidad	Compacidad	Concavidad	Punto de Concavidad	Simetria	Diametro Fractal
45	0.06031	0.03110	0.02031	0.17840	0.05587	0.23850
46	0.12180	0.10440	0.05669	0.18950	0.06870	0.23660
47	0.21900	0.21070	0.09961	0.23100	0.06343	0.98110
48	0.14360	0.09847	0.06158	0.19740	0.06782	0.37040
49	0.10470	0.08259	0.05252	0.17460	0.06177	0.19380
50	0.16860	0.19740	0.10090	0.19070	0.06049	0.62890
51	0.12310	0.12260	0.07340	0.21280	0.06777	0.28710
52	0.07081	0.05253	0.03334	0.16160	0.05684	0.31050
53	0.12670	0.13230	0.08994	0.19170	0.05961	0.72750
54	0.13650	0.12930	0.08123	0.20270	0.06758	0.42260
55	0.20080	0.21350	0.08653	0.19490	0.07292	0.70360
56	0.12620	0.11280	0.06873	0.19050	0.06590	0.42550
57	0.14790	0.12670	0.09029	0.19530	0.06654	0.35770
58	0.18300	0.16920	0.07944	0.19270	0.06487	0.59070
59	0.12800	0.07789	0.05069	0.16620	0.06566	0.27870
60	0.17470	0.23520	0.11960	0.26160	0.06752	1.22300
61	0.08086	0.10260	0.07118	0.19050	0.05265	0.45900
62	0.17910	0.19370	0.14690	0.16340	0.07224	0.51900
63	0.11520	0.12060	0.06470	0.19680	0.05050	0.90480
64	0.09947	0.12040	0.04938	0.20750	0.05636	0.42040
65	0.12060	0.14680	0.08271	0.19530	0.05629	0.54950
66	0.15530	0.17000	0.08815	0.18550	0.06284	0.47680
67	0.15150	0.16370	0.10150	0.18000	0.05641	0.75690
68	0.11410	0.09388	0.05839	0.18790	0.06390	0.28950
69	0.18360	0.20130	0.07798	0.18940	0.06625	0.46340
70	0.17650	0.20710	0.09601	0.19250	0.07692	0.39080
71	0.16490	0.16900	0.08923	0.21570	0.06768	0.42660
72	0.17520	0.21330	0.09479	0.20960	0.07331	0.55200
73	0.06722	0.07293	0.05596	0.21290	0.05025	0.55060
74	0.11000	0.14570	0.08665	0.19660	0.06213	0.71280
75	0.28670	0.42680	0.20120	0.26550	0.06877	1.50900
76	0.08028	0.09271	0.05627	0.19460	0.05044	0.68960
77	0.15890	0.25450	0.11490	0.22020	0.06113	0.49530
78	0.09462	0.07135	0.05933	0.18160	0.05723	0.31170
80	0.12050	0.15230	0.08636	0.23560	0.05635	0.83090
81	0.05761	0.04711	0.02704	0.15850	0.06065	0.23670
82	0.13050	0.15390	0.08624	0.19570	0.06216	1.29600
83	0.11370	0.09447	0.05943	0.18610	0.06248	0.70490
84	0.17000	0.16590	0.07415	0.26780	0.07371	0.31970
85	0.16650	0.18550	0.10540	0.19710	0.06166	0.81130
86	0.11340	0.14980	0.11210	0.17300	0.06147	1.05100
87	0.16660	0.25080	0.12860	0.20270	0.06082	0.73640
88	0.11450	0.13240	0.09702	0.18010	0.05553	0.66420

Pacientes	Suavidad	Compacidad	Concavidad	Punto de Concavidad	Simetria	Diametro Fractal
89	0.09182	0.08422	0.06576	0.18930	0.05534	0.59900
90	0.16120	0.21190	0.11550	0.23750	0.06542	0.85410
91	0.06288	0.05858	0.03438	0.15980	0.05671	0.46970
92	0.15550	0.20320	0.10970	0.19660	0.07069	0.42090
93	0.19140	0.28710	0.18780	0.18000	0.05770	0.83610
94	0.28320	0.24870	0.14960	0.23950	0.07398	0.62980
95	0.08799	0.06593	0.05189	0.16180	0.05549	0.36990
96	0.10520	0.05375	0.03263	0.17270	0.06317	0.20540
97	0.08468	0.08169	0.05814	0.16210	0.05425	0.25770
98	0.24130	0.19810	0.06618	0.23840	0.07542	0.28600
99	0.13530	0.10850	0.04562	0.19430	0.06937	0.40530
100	0.19800	0.16970	0.08878	0.17370	0.06672	0.27960
101	0.12670	0.13850	0.06526	0.18340	0.06877	0.61910
102	0.14280	0.11140	0.06772	0.17670	0.05529	0.43570
103	0.12060	0.11800	0.05980	0.19500	0.06466	0.20920
104	0.20840	0.35230	0.16200	0.22000	0.06229	0.55390
105	0.17680	0.15580	0.09176	0.22510	0.07421	0.56480
106	0.09588	0.07550	0.04079	0.15940	0.05986	0.27110
107	0.15160	0.32010	0.15950	0.16480	0.05525	2.87300
108	0.11460	0.16820	0.06597	0.13080	0.05866	0.52960
109	0.13060	0.11150	0.06462	0.22350	0.06433	0.42070
110	0.15170	0.09901	0.05602	0.21060	0.06916	0.25630
111	0.13060	0.12720	0.08691	0.20940	0.05581	0.95530
112	0.11300	0.11450	0.06637	0.14280	0.05313	0.73920
113	0.15890	0.17690	0.08451	0.19810	0.06613	0.29580
114	0.17990	0.16950	0.06861	0.21230	0.07254	0.30610
115	0.15720	0.19100	0.10900	0.21310	0.06325	0.29590
116	0.10740	0.15540	0.08340	0.14480	0.05592	0.52400
117	0.16820	0.19500	0.12370	0.19090	0.06309	1.05800
118	0.12980	0.14170	0.08811	0.18090	0.05966	0.53660
119	0.15580	0.20490	0.08886	0.19780	0.06000	0.52430
120	0.10410	0.12660	0.08353	0.18130	0.05613	0.30930
121	0.11880	0.13790	0.08591	0.17760	0.05647	0.59590
122	0.12790	0.09789	0.05246	0.19080	0.06130	0.42500
123	0.20630	0.17840	0.11440	0.18930	0.06232	0.84260
124	0.15300	0.16460	0.08722	0.16690	0.06368	0.58790
125	0.22840	0.24480	0.12420	0.23980	0.07596	0.65920
126	0.09398	0.12360	0.09944	0.18480	0.05966	0.38680
127	0.31140	0.31760	0.13770	0.24950	0.08104	1.29200
128	0.16390	0.17510	0.08399	0.20910	0.06650	0.24190
129	0.10880	0.15190	0.09333	0.18140	0.05572	0.39770
130	0.09265	0.13110	0.05591	0.18390	0.05403	0.61480
131	0.12730	0.09697	0.07507	0.21080	0.05464	0.83480

Pacientes	Suavidad	Compacidad	Concavidad	Punto de Concavidad	Simetria	Diametro Fractal
132	0.08995	0.09061	0.06527	0.18670	0.05580	0.42030
133	0.11430	0.13670	0.08646	0.17690	0.05674	1.17200
134	0.19610	0.21950	0.10880	0.17210	0.06194	1.16700
135	0.07027	0.05699	0.04744	0.15380	0.05510	0.42120
136	0.13520	0.16960	0.10430	0.18950	0.05863	0.43520
137	0.18020	0.19480	0.09052	0.18760	0.06684	0.28730
138	0.07338	0.08262	0.05776	0.17970	0.05608	0.26230
139	0.05914	0.02685	0.03515	0.16190	0.06287	0.64500
140	0.16420	0.21970	0.10620	0.17920	0.06552	1.11100
141	0.18380	0.22830	0.12800	0.22490	0.07469	1.07200
142	0.11170	0.11300	0.07950	0.18070	0.05664	0.40410
143	0.11120	0.07333	0.02979	0.15770	0.07097	0.27930
144	0.08564	0.11550	0.07726	0.19280	0.05096	0.59250
145	0.18750	0.25650	0.15040	0.25690	0.06670	0.57020
146	0.13190	0.14780	0.08488	0.19480	0.06277	0.43750
147	0.10560	0.15080	0.09934	0.17270	0.06071	0.81610
148	0.14020	0.10600	0.06090	0.19530	0.06083	0.64220
149	0.12830	0.23080	0.14100	0.17970	0.05506	1.00900
150	0.13390	0.18630	0.11030	0.20820	0.05715	0.62260
151	0.23640	0.29140	0.12420	0.23750	0.07603	0.52040
152	0.23630	0.33680	0.19130	0.19560	0.06121	0.99480
153	0.09769	0.12350	0.06553	0.16470	0.06464	0.65340
154	0.11310	0.09799	0.07785	0.16180	0.05557	0.57810
155	0.16690	0.16410	0.12650	0.18750	0.06020	0.97610
156	0.13120	0.25310	0.12550	0.15210	0.05227	3.36700
157	0.19540	0.24480	0.15010	0.18240	0.06140	1.00800
158	0.10760	0.15270	0.08941	0.15710	0.05478	0.61370
159	0.21540	0.16890	0.06367	0.21960	0.07950	0.21140
160	0.06636	0.08390	0.05271	0.16270	0.05416	0.41570
161	0.13180	0.18560	0.10210	0.19890	0.05884	0.61070
162	0.20870	0.28100	0.15620	0.21620	0.06606	0.62420
163	0.25760	0.31890	0.11980	0.21130	0.07115	0.40300
164	0.13040	0.12010	0.08824	0.19920	0.06069	0.45370
165	0.04605	0.04686	0.02739	0.18520	0.05294	0.46810
166	0.15710	0.15220	0.08481	0.20850	0.06864	1.37000
167	0.11240	0.11170	0.08377	0.17490	0.05822	0.90470
168	0.14890	0.21330	0.12590	0.17240	0.06053	0.43310
169	0.13890	0.15940	0.08744	0.19430	0.06132	0.81910
170	0.11330	0.11260	0.06463	0.16690	0.06544	0.22080
171	0.11090	0.12040	0.05736	0.14670	0.05407	0.51000
172	0.14110	0.12420	0.07569	0.18710	0.05774	0.41250
173	0.11100	0.10070	0.06431	0.17930	0.06281	0.92910
174	0.13140	0.16980	0.08293	0.17130	0.05916	0.38970

Pacientes	Suavidad	Compacidad	Concavidad	Punto de Concavidad	Simetria	Diametro Fractal
175	0.11750	0.15720	0.11550	0.15540	0.05660	0.66430
176	0.09871	0.16550	0.09063	0.16630	0.05391	0.46740
177	0.20040	0.21360	0.10020	0.16960	0.07369	0.92890
178	0.14480	0.22560	0.11940	0.18230	0.06115	0.56590
179	0.07112	0.03649	0.02307	0.18460	0.05325	0.24730
180	0.12890	0.11700	0.07762	0.21160	0.06077	0.75480
181	0.13170	0.14910	0.09183	0.18320	0.06697	0.79230
182	0.16440	0.21880	0.11210	0.18480	0.06222	0.59040
183	0.16810	0.13570	0.06759	0.22750	0.07237	0.47510
184	0.11360	0.17270	0.10710	0.15910	0.05175	1.73000
185	0.18700	0.20300	0.08520	0.18070	0.07083	0.33310
186	0.08597	0.07486	0.04335	0.15610	0.05915	0.38600
187	0.12480	0.15690	0.09451	0.18600	0.05941	0.54490
188	0.13100	0.14110	0.09431	0.18020	0.06188	0.50790
189	0.21060	0.23100	0.14710	0.19910	0.06739	0.99150
190	0.11540	0.14630	0.06139	0.19260	0.05982	0.20270
191	0.11080	0.12750	0.07493	0.18550	0.05593	0.37290
192	0.11050	0.15520	0.07727	0.20600	0.06424	0.55000
193	0.20870	0.25500	0.09429	0.21280	0.07152	0.26020
194	0.22360	0.31740	0.14740	0.21490	0.06879	0.96220
195	0.11590	0.24390	0.13890	0.17260	0.05623	1.17600
196	0.10340	0.14400	0.09791	0.17520	0.05533	0.76550
197	0.10230	0.09251	0.05302	0.15900	0.05648	0.45640
198	0.08318	0.07671	0.05627	0.14330	0.05484	0.61420
199	0.17160	0.21770	0.09718	0.20630	0.06194	0.58540
200	0.17230	0.20530	0.10100	0.17960	0.05715	0.88420
201	0.09175	0.11050	0.06741	0.14240	0.05563	0.45870
202	0.09240	0.06214	0.05598	0.16030	0.05468	0.42360
203	0.22840	0.27020	0.13690	0.23070	0.06308	0.55000
204	0.12670	0.15370	0.07776	0.17700	0.06242	0.46000
205	0.05240	0.03778	0.02876	0.15800	0.05395	0.32920
206	0.09030	0.07268	0.04475	0.18900	0.05690	0.20590
207	0.11970	0.06435	0.08870	0.19710	0.05417	1.34700
208	0.08731	0.09483	0.04286	0.19950	0.05626	0.50160
209	0.20820	0.35790	0.18420	0.25240	0.05848	1.08300
210	0.17390	0.19540	0.12190	0.19810	0.06306	0.79370
211	0.13120	0.11610	0.06403	0.17910	0.07058	0.32230
212	0.19480	0.29790	0.15220	0.17990	0.06340	1.81900
213	0.16990	0.30760	0.15940	0.20990	0.06803	1.46300
214	0.06271	0.06151	0.03938	0.19930	0.06554	0.37080
215	0.10770	0.15990	0.08705	0.16200	0.05731	0.82150
216	0.11670	0.13550	0.08397	0.16000	0.05461	0.52990
217	0.12160	0.11830	0.07227	0.21260	0.06105	0.53050

Pacientes	Suavidad	Compacidad	Concavidad	Punto de Concavidad	Simetria	Diametro Fractal
218	0.07572	0.08134	0.04838	0.17910	0.06214	0.45400
219	0.06237	0.06137	0.05157	0.15270	0.05223	0.50330
220	0.17770	0.21380	0.11160	0.19240	0.06389	0.49700
221	0.21120	0.17290	0.09465	0.20790	0.07496	0.34050
222	0.05998	0.04656	0.03905	0.16520	0.05182	0.36330
223	0.05998	0.04656	0.03905	0.16520	0.05182	0.36330
224	0.16220	0.26670	0.13630	0.16050	0.06261	0.65150
225	0.12330	0.13650	0.07268	0.21500	0.06098	0.51420
226	0.09162	0.06862	0.06367	0.17280	0.05262	1.37400
227	0.10620	0.13750	0.06500	0.17350	0.06105	0.32350
228	0.06655	0.06583	0.05068	0.17930	0.05392	0.61010
229	0.11310	0.10120	0.04989	0.18900	0.06035	0.60520
230	0.12140	0.11570	0.05378	0.20740	0.06949	0.49180
231	0.10050	0.15460	0.10050	0.16310	0.05502	1.32000
232	0.14800	0.08524	0.06253	0.18900	0.06396	0.71470
233	0.10850	0.16870	0.13700	0.15390	0.04860	1.55800
234	0.11940	0.18100	0.11500	0.18650	0.05276	0.79060
235	0.09319	0.08891	0.04750	0.15170	0.05432	0.54980
236	0.09391	0.10560	0.05746	0.15850	0.05405	0.17060
237	0.23070	0.32860	0.12800	0.27060	0.06065	0.51210
238	0.15190	0.22490	0.14550	0.17710	0.05882	1.25800
239	0.23340	0.39370	0.18070	0.17560	0.06358	0.79530
240	0.08977	0.07115	0.05711	0.18130	0.05899	0.53200
241	0.11980	0.13790	0.07762	0.17290	0.06281	0.39760
242	0.13180	0.17470	0.08273	0.17090	0.05953	0.43820
243	0.14820	0.21950	0.13010	0.20320	0.05875	1.27700
244	0.09576	0.16180	0.10480	0.18010	0.05508	0.81340
245	0.07867	0.09958	0.09479	0.15350	0.04968	0.81730
246	0.10180	0.09318	0.05451	0.18800	0.05692	0.24550
247	0.10110	0.10500	0.08129	0.21360	0.05323	0.30230
248	0.16590	0.23640	0.13710	0.19370	0.05372	0.82410
249	0.06273	0.06314	0.05896	0.20290	0.04655	0.90140
250	0.06674	0.09007	0.06465	0.17360	0.04780	0.62240
251	0.11670	0.09041	0.07284	0.18940	0.06002	0.31320
252	0.09069	0.11200	0.12350	0.15260	0.04905	1.61700

Bibliografía

[1] Armitage P. *Statistical methods in medical research*, Blackwell Scientific Publivation. 2002

[2] Collet, D.S. *Modelling Survival Data in Medical Research*, Chapman and Hall. Londres. 2003

[3] Cox, D.R.,Oakes, D. *Analysis of Survival Data*, Chapman and Hall. Grate Britain. 1984

[4] Deshpande, Jayant V. *Life-time Data; Statistical Models and Methods.*, Series On Quality, Reliability and Engineering Statistics, Vol. 11. 2005

[5] Dobson Annette J. *An introduction to generalized linear models*, Chapman and Hall. Londres. 2008

[6] Hougaard, Philips. *Analysis of Multivariate Survival Data*, Springer. 2000

[7] Hougaard, Philips. *Fundamentals of Survival Data. Biometrics*, Springer. 1999

[8] Kalbfleisch, J.D., Prentice, R.L. *The Statistical Analysis of Failure Time Data*, John Wiley and Sons. New York. 2002

[9] Korn, Edward L. *Analysis of health surveys*, John Wiley and Sons. New York. 1999

[10] Kleinbaum, D.G. *Survival analysis: a selft-learning text*, Springer-Verlag. New York. 1996

[11] Klein J.P., Moeschberger M.L. *Survival analysis. Thechniques for censured and truncated date*, Springer-Verlag. New York. 2003

[12] Lachin Jhon M. *Biostadistical Methods. The Assessment of relative risks*, John Wiley and Sons. New York. 2000

[13] Lawless, J.F. *Statistical Models and Methods for Lifetime Data*, John Wiley and Sons. New York. 2003

[14] Lee T. Elisa. *Statistical Methods for survival data analysis*, John While and Sons. New York. 2003

[15] Mahesh K.B. Parmar and Machin *Survival analysis. A practical approach*, John Wiley and Sons. New York. 1995

[16] Maller Ross. *Survival analysis with long-term survivors*, John Wiley and Sons. New York. 1996

[17] Miller, R.G. *Survival Analysis*, John Wiley and Sons. New York. 1981

[18] Mood, A.*Introduction to The Theory of Statistics*, Mc. Graw-Hill. Singapure. 1974

[19] Smith Peter J. *Analysis of failure and survival data*, Chapman and Hall. Londres. 2002

[20] Therneau Terry M. *Modelling survival data: extending the cox model*, Springer. New York 2000