Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias Físico Matemáticas

Punto de cambio a través de ensayo y error del Modelo de Cox

Tesis presentada al Colegio de Matemáticas para la obtención del título de:

Licenciada en Actuaría

Presenta:
Anayeli Cocoletzi Conde

Director de Tesis:
Dr. Francisco Solano Tajonar Sanabria

Puebla, Pue.
Noviembre, 2021

A mi madre, Obdu, a mi sobrina Nat,
a mis hermanos Boba, Nachito y Jo.
En memoria de mi padre, Meli.
Los amo...!

Agradecimientos

A Dios por haberme permitido vivir hasta este día, por darme la capacidad física e intelectual para culminar con éxito una etapa más, por permitirme llegar hasta donde los demás no creían que podía hacerlo.

A mis padres por ser mis principales motivadores y cómplices, gracias a ellos por cada día confiar y creer en mí, por seguir mis ideas locas. Gracias mami por tus consejos, por tu apoyo incondicional, por llorar conmigo en los malos momentos y después de todo decirme que las cosas iban a estar bien. Mami te amo. A mi padre, donde quiera que estes papi, gracias por preocuparte cada día por mí, por cada palabra de aliento, por enseñarme a valorar las cosas y por ser mi ejemplo a seguir en todo momento, papi este logro no hubiese sido posible sin ti, no estás aquí pero estaré eternamente agradecida con dios por haber tenido un padre como tú. Papi siempre te amaré.

A mis hermanos y sobrinos, por compartir alegrías y malas rachas, por permitirme formar parte de su vida, por escucharme y apoyarme en cada paso. Todos ustedes ocupan un lugar especial en mi corazón.

A mi Asesor de Tesis, el Dr. Tajonar por haberme permitido acudir a él, por transmitirme parte de sus grandes conocimientos, así también por tenerme toda la paciencia del mundo para terminar mi trabajo y por dedicarle de su valioso tiempo a la revisión del mismo. Infinitas gracias.

A la Mtra. Zavala, al Dr. Velasco, a la Dra. Reyes por su tiempo para echarle un vistazo a mi trabajo de tesis y por darme sus comentarios para mejorar este escrito, les agradezco con todo mi ser.

A todas aquellas personas que hoy están aquí y a las que un día estuvieron, gracias por formar parte de mi vida y por llenarme de buenos recuerdos, sin importar donde estén, muchas gracias de todo corazón.

Índice general

Introducción

En la vida cotidiana ocurren eventos aleatorios de los cuales posiblemente no se pueden definir las características relacionadas a lo sucedido de manera sencilla, más aún, no se puede definir a simple vista una función que permita conocer el riesgo de que ocurra dicho evento, para estos casos existe el análisis de supervivencia, una área bastante conocida que permite modelar y analizar conjuntos de datos relacionados con una variable de salida.

El análisis de supervivencia es una rama de la estadística inferencial cuya teoría va de la mano con la teoría de la confiabilidad, donde el objetivo principal es determinar la probabilidad de que un evento ocurra durante o al finalizar un determinado periodo de tiempo, para lograr este objetivo el análisis de supervivencia hace uso de las herramientas de la estadística inferencial, las cuales nos permiten describir, medir y relacionar a un conjunto de características o atributos con una variable de salida cuya naturaleza es aleatoria, es decir, si definimos a T como la variable de interés siendo ésta de tipo continuo o discreto, para decribir su comportamiento el análisis de supervivencia hará uso de las funciones básicas como lo son la función de riesgo, función de densidad de probabilidad y función de supervivencia.

Como en todas las áreas de investigación existen casos donde los conjuntos de datos ajustan excelente con los modelos básicos, sin embargo, también existen aquellas excepciones donde la naturaleza de los datos no lo permite, dificultando el proceso de ajuste, por ello dentro del área de análisis de supervivencia existen diversos modelos que van desde los más sencillos, como lo son los modelos paramétricos hasta los no paramétricos. Dentro de estos últimos se encuentra el modelo de Cox, también denominado modelo de Riesgos Proporcionales propuesto por David Cox en 1985 cuyo enfoque nos permite analizar tiempos de falla incluyendo el efecto de un conjunto de covariables que pueden ser constantes durante el periodo de estudio u observación o bien cambiar de estado en algún punto dentro del periodo de tiempo.

De manera que, el objetivo de este trabajo de tesis es estudiar el famoso modelo de Cox con un punto de cambio para aplicarlo a un conjunto de datos financieros, en el cual el análisis de supervivencia permita definir una función de riesgo donde el evento de interés esta en función de una variable de salida y de un conjunto predefinido de variables explicativas denominadas

covariables. El conjunto de datos contiene una muestra de trabajadores afiliados a AFORE, la cual a su vez contiene 8 variables que describen a un trabajador que es o fue afiliado a la empresa, lo que da paso a que el objetivo del estudio sea determinar una función de riesgo que contemple el efecto que surge de considerar las características que describen a cada trabajador, esto con el fin de conocer si estas influyen en que el trabajador se cambie de AFORE o no, además se buscará determinar si existe un punto de cambio en el que la función de riesgo varíe de acuerdo al estado de una característica.

Para lograr el cometido la estructura de esta tesis con respecto a los temas contenidos sigue el siguiente orden:

En el Primer Capítulo se hace una introducción al área de análisis de supervivencia, en él se describen los conceptos básicos, así como el planteamiento y demostración de los resultados más importantes que se pueden hallar dentro del área de supervivencia. Adicional a lo anterior se enlistan las distribuciones más conocidas que permiten modelar una función de riesgo con sus respectivos parámetros.

En el Segundo Capítulo se describe el modelo de Cox, la versión clásica que supone riesgos proporcionales en las covariables asociadas al evento de interés, además, se enlistan sus características principales, ventajas y desventajas, unido a esto también se plantea el método de máxima verosimilitud parcial con el fin de entender cómo es que se estiman los parámetros que acompañan a las covariables cuando se intenta medir el riesgo de falla de los elementos de estudio.

En el Tercer Capítulo se plantea en primer lugar el modelo exponencial con un punto de cambio, enseguida se plantea la extensión del modelo de Cox con un punto de cambio, este capítulo es de suma importancia ya que el entendimiento de éste ayudará a la correcta aplicación del modelo en el conjunto de datos.

En el Cuarto Capítulo se hace la aplicación del modelo de Cox con un punto de cambio, sin embargo, antes de generar el modelo el capítulo comienza con la aplicación del modelo de Cox clásico que considera riesgos proporcionales para analizar el comportamiento de los datos, hasta entonces se generará la extensión del modelo, donde la determinación del punto de cambio se hará bajo el método de ensayo y error.

Al concluir los capítulos anteriores y para finalizar este escrito se agregan las conclusiones obtenidas de la aplicación.

Capítulo 1

Conceptos preliminares del análisis de supervivencia

En este capítulo se presentan los conceptos básicos del análisis de supervivencia, la cual es una línea de investigación de la estadística inferencial que permite entender métodos y procedimientos para inferir acerca del comportamiento de una población. En este análisis, la población de interés concentra a un grupo o grupos de individuos donde cada uno tiene un tiempo de falla, el cual se puede considerar como variable aleatoria, digamos T, la cual representa el tiempo que transcurre desde que se inicia un estudio hasta que ocurre un evento que se denomina falla, es decir, el elemento llega a un evento final donde deja de presentar características generales de la población.

Entonces, si consideramos una muestra de n elementos, el objetivo principal será incorporar toda la información disponible para describir a la muestra, en otras palabras, se busca conocer a la muestra mediante las funciones de medida fundamentales que brinda la teoría de supervivencia como son: la función de densidad de probabilidad, la función de distribución acumulada, la función de supervivencia, así como, la tasa de riesgo que a continuación serán definidas junto con las relaciones y propiedades que las caracterizan. Unido a esto también se mencionan las principales distribuciones que ajustan mejor a un conjunto de tiempos de falla de una muestra de individuos, pues existen muchas distribuciones pero no todas describen de manera correcta al conjunto de datos que se tiene, ya sea porque el modelo no es el adecuado o porque como en todo conjunto de datos, la información se puede encontrar incompleta, es decir, los datos han sido censurados. La censura se presenta cuando algunos datos son desconocidos, este concepto también será discutido en este capítulo junto con sus principales tipos de censura.

1.1. Tiempo de falla

El concepto de tiempo de falla es y ha sido utilizado en el campo de la fiabilidad para distintas situaciones, por ello es conveniente definir primero qué se entiende por falla, para no tener ambigüedades con la definición de tiempo de falla.

Definición 1.1.1 *Se entiende por falla al punto donde el elemento de estudio presenta el cambio de estado.*

De manera que, una falla puede interpretarse entonces como el punto de desequilibrio del elemento de interés , por ejemplo, puede ser la respuesta de aplicación de un tratamiento, la recaída de un paciente por el desarrollo de una enfermedad o bien el punto donde se otorga la libertad a una persona, etc.

Así, si consideramos la definición anterior, el tiempo de falla puede definirse de la siguiente manera.

Definición 1.1.2 *El tiempo de falla es el tiempo que transcurre a partir de un evento inicial hasta que ocurre el evento puntual que involucra un final para cada elemento de la muestra. El valor del tiempo de falla por relacionarse con el tiempo es estrictamente mayor o igual que cero.*

Un tiempo de falla puede considerar diferentes situaciones, dependiendo del campo de estudio donde se encuentre el investigador, algunos ejemplos son los siguientes,

- Tiempo de vida útil del algún dispositivo o sistema electrónico (Campo de la confiabilidad)
- Tiempo que puede durar un matrimonio (Sociología)
- Tiempo que dura un persona en desarrollar una cierta enfermedad, hasta que llega a morir o hasta que reacciona a algún tratamiento (Medicina)

El tiempo de falla se considera un factor importante en la toma de decisiones, por ello se requiere que su estudio tenga una determinación y comprensión precisa. Por ejemplo, digamos que una cierta empresa dedicada al ensamblaje de piezas tiene maquinaria un poco desgastada, por lo cual está decidida a cambiar de maquinaria siempre y cuando sus tiempos de ensamblaje excedan un cierto número de horas, por ello determinar los tiempos de manera

adecuada le permitirá a la empresa tomar la mejor decisión acerca de su maquinaria, de lo contrario sólo podría perder tiempo y dinero.

En consecuencia a lo anterior, se definen tres requisitos para poder determinar el tiempo de falla de manera correcta.

1.1.1. Requisitos para determinar el tiempo de falla

- Origen: es el punto donde se inicia el estudio u observación de la muestra. Por ejemplo, si se considera un grupo de niños, de los cuales se desea estudiar el desarrollo de bichos a partir de la aplicación de un nuevo medicamento. Es claro que el punto de origen puede variar en cada niño, pues la aplicación del medicamento puede ser en diferentes edades.
- Escala para medir el tiempo: se refiere a la unidad de medida con la que se tomará la longitud del tiempo que transcurre hasta que llega la falla, por ejemplo, si se considera el caso de la empresa que desea renovar su maquinaria, la escala adecuada serían las horas, de esta manera se podría decir que la escala depende directamente de la población de estudio.
- Concepción del evento de falla: considera el hecho que lleva a cada elemento de la muestra a un estado que se puede considerar inactivo, para este requisito se requiere que se tenga claro el significado del evento ya que una mala interpretación podría definir un tiempo de falla de manera incorrecta. Por ejemplo, si se sigue considerando a la compañía que debe decidir entre cambiar su maquinaria o no hacerlo, la falla ocurrirá cuando su ensamblaje exceda un cierto número de horas fijado por la misma compañía o de manera más general exceda un tiempo t.

Observación 1.1.3 *Cuando el centro de estudio del análisis de supervivencia se encuentra en una muestra de individuos, el tiempo de falla es denominado también como "Tiempo de vida".*

Ya que se tiene claro que es un tiempo de falla, no es difícil notar que el tiempo que comprende la definición se encuentra sujeto a variaciones aleatorias, dicho de otro modo se considera una variable aleatoria, digamos T_i, para representar el tiempo de falla de cada elemento i de la muestra.

Así, si consideramos una muestra de una población homogénea de n individuos, donde a cada uno le corresponde un tiempo de falla, tendríamos n variables aleatorias que podrían seguir un cierto patrón, es decir, sería conocer el comportamiento de la población a través de las funciones básicas con las que se describen las variables aleatorias.

Observación 1.1.4 *Debido a que la población es de individuos, se considera que las variables aleatorias son continuas ya que las unidades de medida se encuentran relacionadas con las unidades de medida naturales como los días, meses y años.*

A continuación se presentan las funciones básicas con las que habitualmente se representa el comportamiento de una variable aleatoria. Para lo anterior formalmente consideremos una población homogénea de individuos, y sea T una variable aleatoria continua que representa el tiempo de falla de cualquier individuo de la población.

1.2. Función de densidad de probabilidad

Cuando hablamos de la función de densidad de probabilidad en el análisis de supervivencia, no sólo nos referimos a una función que proporciona una probabilidad en un intervalo, sino también a aquella que nos permite obtener de manera única la función de distribución acumulada de una variable aleatoria con la cual conoceremos más a detalle del comportamiento de dicha variable.

Definición 1.2.1 *Sea T una variable aleatoria continua. Entonces para T existe una función denotada por $f(t)$, denominada función de densidad de probabilidad, que satisface las siguientes propiedades [6],*

- *Para toda t en el recorrido de la variable $(-\infty, +\infty)$ se tiene que*

$$f(t) \geq 0. \tag{1.1}$$

- $$\int_{-\infty}^{+\infty} f(t)\, dt = 1. \tag{1.2}$$

- *Para cualquier t_1, $t_2 \in (-\infty, +\infty)$ tal que $t_1 \leq t_2$, se tiene que,*

$$P\,[\ t_1 \leq T \leq t_2\ ] = \int_{t_1}^{t_2} f(x)\, dx. \tag{1.3}$$

1.3. Función de distribución acumulada

Definición 1.3.1 *Denotada por $F(t)$, se define a la función de distribución acumulada de la variable aleatoria T como la probabilidad de que el elemento falle antes o hasta el tiempo*

t, es decir,

$$\begin{aligned} F(t) &= P\,[\ Un\ elemento\ falle\ antes\ o\ hasta\ el\ tiempo\ t\ ] \\ &= P\,[\ T \leq t\ ]\,. \end{aligned} \tag{1.4}$$

A continuación, se presentan las propiedades que cumple la función de distribución acumulada, además de, la demostración de las dos primeras propiedades.

Propiedades 1.3.2 *Sea $F(t)$ la función de distribución acumulada de la variable aleatoria T, entonces, $F(t)$ satisface lo siguiente [6],*

1. *La función $F(t)$ es no decreciente, es decir, si $t_1 \leq t_2$ entonces $F(t_1) \leq F(t_2)$.*

2. $\lim_{t\to-\infty} F(t) = 0$ *y* $\lim_{t\to+\infty} F(t) = 1$.

3. *La función $F(t)$ es continua a la derecha.*

Demostración de a):

Sea $A = T \leq t_1$ y $B = T \leq t_2$ eventos cualquiera. Supongamos además que $t_1 \leq t_2$. Entonces de la teoría de conjuntos se sigue que A está contenido en B, o que A es un subconjunto de B. Por lo que, $F(t_1) \leq F(t_2)$. ■

Demostración de b):

Para la primera expresión se puede abordar de la siguiente manera,

$$\lim_{t\to-\infty} F(t) = \lim_{t\to-\infty} \int_{-\infty}^{t} f(x)\,dx = 0.$$

Mientras que para la segunda expresión se tiene que,

$$\lim_{t\to+\infty} F(t) = \lim_{t\to+\infty} \int_{-\infty}^{t} f(x)\,dx = 1.$$

Con lo cual queda demostrado. ■

Otro resultado de importancia significativa para la función de distribución de acumulación, es la relación que existe entre ésta y la función de densidad de probabilidad.

Proposición 1.3.3 *Sea $f(t)$ la función de densidad de probabilidad de la variable aleatoria T, entonces,*

$$f(t) = \frac{dF(t)}{dt}. \tag{1.5}$$

Demostración:

De la Definición 1.2.1 se interpreta a $f(t)$ como una medida instantánea , la cual se define como el límite de probabilidad de que un individuo falle en el intervalo $(t, t + \triangle t)$, es decir,

$$\begin{aligned} f(t) &= \frac{\lim_{\triangle t \to 0} P\,[\ Un\ individuo\ falle\ en\ (t, t+\triangle t)\ ]}{\triangle t} \\ &= \lim_{\triangle t \to 0} \frac{P\,[\ t < T \leq t + \triangle t\ ]}{\triangle t} \\ &= \lim_{\triangle t \to 0} \frac{P\,[\ T \leq t + \triangle t\ ] - P\,[\ T \leq t\ ]}{\triangle t} \\ &= \lim_{\triangle t \to 0} \frac{F(t + \triangle t) - F(t)}{\triangle t} \\ &= \frac{dF(t)}{dt}. \end{aligned}$$

Con lo cual queda demostrado lo requerido. ■

Además, agregamos que la función de densidad de probabilidad se considera también como una función de densidad incondicional de falla [2].

1.4. Función de supervivencia

Definición 1.4.1 *Denotada por $S(t)$, se define a la función de supervivencia de la variable aleatoria T como la probabilidad de que un elemento falle después del tiempo t, así,*

$$\begin{aligned} S(t) &= P\,[\ Un\ elemento\ sobreviva\ después\ del\ tiempo\ t\ ] \\ &= P\,[\ T > t\ ] = \int_t^{\infty} f(x)\,dx. \end{aligned} \tag{1.6}$$

Otra forma de escribir esta probabilidad es utilizando la función de distribución, esto es

$$\begin{aligned} S(t) &= P[\ Un\ \ elemento\ \ sobreviva\ \ después\ \ del\ \ tiempo\ \ t\ ] \\ &= 1 - P[\ Un\ \ elemento\ \ falle\ \ antes\ \ o\ \ hasta\ \ el\ \ tiempo\ \ t\ ] \\ &= 1 - P[\ T \leq t\ ] \\ &= 1 - F(t). \end{aligned} \tag{1.7}$$

De esta forma se tiene que $S(t) = 1 - F(t)$.

Para la función de supervivencia también existen resultados análogos a los que se mencionaron en la parte correspondiente a la función de distribución.

Propiedades 1.4.2 *Sea $S(t)$ la función de supervivencia de la variable aleatoria T, entonces se satisface lo siguiente [6],*

1. *La función $S(t)$ es decreciente, se tiene que si $t_1 \leq t_2$ entonces $S(t_1) \geqslant S(t_2)$.*

2. $\lim_{t\to-\infty} S(t) = 1 \quad y \quad \lim_{t\to+\infty} S(t) = 0.$

Demostración de a):

Como la demostración de a) es análoga a la demostración de las propiedades de la función de distribución, se prosigue a demostrar la propiedad b).

Demostración de b):

Eligiendo a la primera expresión, la demostración queda de la siguiente manera,

$$\begin{aligned} \lim_{t\to-\infty} S(t) &= S(-\infty) = 1 - F(-\infty) \\ &= 1 - \lim_{t\to-\infty} \int_{-\infty}^{t} f(x)\, dx = 1. \end{aligned}$$

Mientras que para la segunda expresión se tiene,

$$\begin{aligned} \lim_{t\to\infty} S(t) &= S(\infty) = 1 - F(\infty) \\ &= 1 - \lim_{t\to\infty} \int_{-\infty}^{t} f(x)\, dx = 0. \end{aligned}$$

Con lo cual queda demostrado. ■

Así, como la función de distribución acumulada se relaciona con la función de densidad de probabilidad, para la función de supervivencia se obtiene un resultado similar.

Proposición 1.4.3 *Sea $f(t)$ la función de densidad de probabilidad de la variable aleatoria T, entonces,*

$$f(t) = -\frac{dS(T)}{dt}. \tag{1.8}$$

Demostración:

Para que esta proposición sea más fácil de probar será conveniente recurrir al resultado que se obtuvo en la Proposición 1.3.3, de manera que,

$$\begin{aligned} f(t) &= \frac{\text{ĺím}_{\triangle t \to 0} P\,[\ Un\ elemento\ falle\ en\ (t, t+\triangle t)\ ]}{\triangle t} \\ &= \frac{dF(t)}{dt} = \frac{d(1-S(t))}{dt} \\ &= \frac{d(1)}{dt} - \frac{dS(t)}{dt} \\ &= -\frac{dS(t)}{dt}. \end{aligned}$$

Que es lo que se quería demostrar. ■

1.5. Función de riesgo

Dentro del análisis de supervivencia la función de riesgo juega papel importante ya que con ella se determina la distribución de probabilidad que puede tener un conjunto de datos de tiempos de falla, los cuales sirven para determinar o pronosticar la probabilidad de que ocurra algún evento que afecte a la población, esta función es conocida también como,

- Función instantánea de falla.
- Fuerza de mortalidad.
- Tasa de falla condicional.
- Tasa de mortalidad condicional.

Definición 1.5.1 *La función de riesgo es denotada por $h(t)$ y se define como la probabilidad de falla durante un intervalo de tiempo infinitesimal $(t, t+\triangle t)$, es decir, el elemento fallará en el intervalo de tiempo $(t, t+\triangle t)$ dado que ya ha sobrevivido hasta el tiempo t [3], es decir,*

$$h(t) = \frac{1}{\triangle t}\text{ĺím}_{\triangle t \to 0} P\,[\ Un\ individuo\ falle\ en\ (t, t+\triangle t)\ \mid\ que\ ya\ ha\ sobrevivido\ hasta\ t\ ]. \tag{1.9}$$

Seguido de la igualdad anterior, entonces se puede interpretar a la función de riesgo como una densidad condicional [2],

$$Densidad\ condicional = \frac{Densidad\ incondicional}{Probabilidad\ de\ que\ sobreviva\ hasta\ t}.$$

Una vez que se tiene la definición de la función de riesgo, se definen las siguientes relaciones entre está y la función de supervivencia.

Lema 1.5.2 *Para toda $t \in [0, \infty)$, tenemos que [7],*

$$h(t) = \frac{f(t)}{S(t)}. \tag{1.10}$$

Demostración:

Usando la Definición 1.5.1, se tiene que,

$$\begin{aligned} h(t) &= \lim_{\triangle t \to 0} \frac{P\,[\ t \leq T < t + \triangle t \mid T \geq t\ ]}{\triangle t} \\ &= \lim_{\triangle t \to 0} \frac{P\,[\ t \leq T < t + \triangle t\ ]}{P\,[\ T \geq t\ ]\,\triangle t} \\ &= \lim_{\triangle t \to 0} \frac{F(t + \triangle t) - F(t)}{S(t)\triangle t} \\ &= \frac{1}{S(t)} \lim_{\triangle t \to 0} \frac{F(t + \triangle t) - F(t)}{\triangle t} \\ &= \frac{f(t)}{S(t)}. \blacksquare \end{aligned}$$

Teorema 1.5.3 *Sea $h(t)$ la función de riesgo y $S(t)$ la función de supervivencia de la variable aleatoria T, entonces las siguientes relaciones son válidas [7],*

a)

$$h(t) = -\frac{d \ln(S(t))}{dt}. \tag{1.11}$$

b)

$$S(t) = \exp - \left\{ \int_0^t h(x)\, dx \right\}. \tag{1.12}$$

Demostración de a):

Usando el Lema 1.5.2, se tiene que,

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} = \frac{1}{S(t)} \frac{dF(t) - 1}{dt} \\ &= -\frac{1}{S(t)} \frac{dS(t)}{dt} = -\frac{d \ln(S(t))}{dt}. \end{aligned}$$

Demostración de b):

Para la demostración de este inciso partiremos del inciso a), y despejando obtenemos que,

$$-\ln(S(t)) = \int_0^t h(x)\, dx$$

$$S(t) = \exp - \left\{ \int_0^t h(x)\, dx \right\}. \blacksquare$$

1.6. Función de riesgo acumulada

En consecuencia a la importancia de la función de riesgo que se mencionó en la sección anterior, la función de riesgo acumulada también tiene un rol importante dentro del análisis de supervivencia. Ésta medida se interpreta como la probabilidad de que un elemento falle en el instante t dado que ha sobrevivido a un $\triangle t$ más de tiempo.

Definición 1.6.1 *Denotada por $H(t)$, se define a la función de riesgo acumulada como,*

$$H(t) = \int_0^t h(x)\, dx. \tag{1.13}$$

Proposición 1.6.2 *Sea $S(t)$ la función de supervivencia de la variable aleatoria T, entonces,*

$$H(t) = -\ln(S(t)). \tag{1.14}$$

Demostración:

Para ésta demostración recurrimos al inciso b) del Teorema 1.5.3. Entonces tenemos que,

$$S(t) = \exp - \left\{ \int_0^t h(x)\, dx \right\}$$

$$\ln(S(t)) = - \int_0^t h(x)\, dx$$

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

1.7. Algunas distribuciones especiales

Uno de los problemas importantes que se tiene en el análisis de supervivencia es determinar si un conjunto de datos de tiempo de falla se ajusta con un modelo paramétrico, semi paramétrico o no paramétrico. Si optamos por un modelo paramétrico es necesario conocer cuáles son las distribuciones más usuales para modelar un conjunto de datos de tiempos de falla, ya que no es suficiente sólo con ajustar el modelo sino también es importante conocer cuáles son las características que permiten tener una mejor compresión acerca del comportamiento del conjunto de datos. Por ello, a continuación se describen las diversas distribuciones que mejor ayudan para ajustar datos que contienen tiempos de falla, junto con algunas de sus características.

1.7.1. Distribución Exponencial

Dentro del análisis de supervivencia el modelo exponencial desempeña un papel importante no sólo por su facilidad de modelación sino también por ser una distribución continua que sólo toma valores positivos. Su aplicación se usa para describir fenómenos de diferentes campos, sin embargo, el área principal de enfoque es la teoría de la confiabilidad.

Definición 1.7.1 *Sea T una variable aleatoria continúa, se dice que T sigue una distribución exponencial con parámetro $\lambda > 0$, si T no toma valores negativos y su función de densidad está dada de la siguiente manera,*

$$f(t) = \lambda \ \exp\{-\lambda \ t \ \}, \quad \textit{con } \lambda > 0. \tag{1.15}$$

Propiedades 1.7.2 *Propiedades de la distribución exponencial*

- *Su función de distribución acumulada está dada por*

$$\begin{aligned} F(t) &= \int_0^t f(x)\,dx \\ &= \int_0^t \ \lambda\ \exp\{-\ \lambda t\ \}\ dx \\ &= 1 - \exp\{\ -\lambda t\ \}\,. \end{aligned} \tag{1.16}$$

- *Mientras que su función de supervivencia puede ser determinada de manera inmediata, una vez que se tiene la expresión anterior*

$$\begin{aligned} S(t) = 1 - F(t) &= 1 - [1 - \exp\{-\lambda t\}] \\ &= \exp\{-\lambda t\}\,. \end{aligned} \tag{1.17}$$

- *Su tasa de riesgo es constante, es decir, la probabilidad de que falle hoy y la probabilidad de que falle dentro de un tiempo t es la misma,*

$$h(t) = \lambda, \quad para \quad todo \quad t. \tag{1.18}$$

Proposición 1.7.3 *Si T es una variable aleatoria continua que se distribuye de manera exponencial, y además se consideran $t_0,\ t > 0$, entonces se satisface la siguiente igualdad,*

$$P\,[\ T > t + t_0 \mid T > t_0\ ] = P\,[\ T > t\ ]\,. \tag{1.19}$$

En otras palabras, la probabilidad condicional es igual a la probabilidad incondicional.

Demostración:

Para cualquier $t > t_0$ se tiene que,

$$\begin{aligned} P\,[\ T > t + t_0 \mid T > t_0\ ] &= \frac{P\,[\ T > t + t_0\ ]}{P\,[\ T > t_0\ ]} \\ &= \frac{\exp\{\ -\lambda(t + t_0)\ \}}{\exp\{\ -\lambda t_0\ \}} \\ &= \exp\{\ -\lambda t\ \} = P\,[\ T > t\ ]\,.\blacksquare \end{aligned}$$

El resultado anterior, es muy importante debido a que con él queda demostrada la propiedad que caracteriza a la distribución exponencial, a esta se le denomina "pérdida de memoria" y significa que la tasa de riesgo para un individuo no depende de la edad que tenga, la tasa de riesgo se mantendrá constante. En otras palabras, mientras el objeto siga funcionando éste se considera como nuevo.

1.7.2. Distribución Weibull

La distribución Weibull es ampliamente utilizada después de la distribución exponencial y además, es una de las que mejor se ajusta a datos de tiempo de vida [3]. Tiene una aplicación más general con respecto a la distribución exponencial, ya que a diferencia ésta no considera una tasa de riesgo constante y cuenta con dos parámetros denominados "parámetro de forma" y "parámetro de escala".

Definición 1.7.4 *Sea T una variable aleatoria continua, se dice que T sigue una distribución Weibull, si su función de densidad es de la forma [3],*

$$f(t) = \alpha\ \beta\ ^{\alpha} t\ ^{\alpha-1} \exp\left\{-(\beta t)\alpha\right\}, \quad \textit{con}\ \alpha,\ \beta > 0. \tag{1.20}$$

Donde α es considerado el parámetro de forma y β el parámetro de escala.

Propiedades 1.7.5 *Propiedades de la distribución Weibull*

- *Función de distribución*
$$F(t) = 1 - \exp\left\{-(\beta t)\alpha\right\}. \tag{1.21}$$
- *Función de supervivencia*
$$S(t) = \exp\left\{-(\beta t)\alpha\right\}. \tag{1.22}$$
- *Función de riesgo*
$$h(t) = \alpha\beta(\beta\alpha)^{\alpha-1}. \tag{1.23}$$

Observación 1.7.6 *Si el parámetro de forma, en este caso α es igualado a 1, se tiene que la variable aleatoria T sigue un modelo exponencial con parámetros $(\beta, \alpha = 1)$.*

De esta manera, si $T \sim Weibull(\beta, \alpha = 1)$, entonces su función de densidad es de la forma,

$$\begin{aligned} f(t) &= \alpha\ \beta\ ^{\alpha}\ t\ ^{\alpha-1}\ \exp\{\ -(\beta t)\alpha\ \} \\ &= \beta\ \exp\{\ -\beta t\ \}. \end{aligned}$$

Donde claramente la función de densidad es igual a la de un modelo exponencial con parámetro β, es decir, el modelo exponencial es un caso particular del modelo Weibull.

Proposición 1.7.7 *Sea T una variable aleatoria que sigue un modelo* $\exp(\beta = 1)$. *Entonces, si se hace una transformación de la forma*

$$X = (aT)^{\frac{1}{\alpha}},$$

con a una constante mayor que cero, se obtiene que X se distribuye de manera Weibull con parámetros $\left(\frac{1}{\alpha}, \alpha\right)$.

Demostración:

Como

$$X = (aT)^{\frac{1}{\alpha}}.$$

Despejando obtenemos que

$$T = \frac{X^{\alpha}}{a}.$$

Así,

$$F(T) = F\left(\frac{X^{\alpha}}{a}\right) = 1 - \exp\left\{\frac{X^{\alpha}}{a}\right\}.$$

Con lo cual queda demostrado, porque la última expresión representa la función de distribución que sigue un modelo Weibull. ■

1.7.3. Distribución Gamma

Antes de enunciar a ésta distribución es importante definir cuál es la función Gamma pues ésta función desempeña un rol importante tanto en la probabilidad como en muchas otras áreas de las matemáticas.

Definición 1.7.8 *La función Gamma denotada por* Γ*, se define como,*

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp\{-x\}\, dx. \tag{1.24}$$

Donde la integral impropia está definida si $\alpha > 0$ *[6].*

Observación 1.7.9 *Al desarrollar el resultado anterior se llega a que la integral impropia existe y que converge siempre que* $\alpha > 0$*. Es decir, si se integra a la expresión anterior por partes, definiendo las siguientes variables auxiliares como siguen,* $\exp\{-x\}\, dx = dv$ *y* $x^{\alpha-1} = u$ *se llega a [6],*

$$\begin{aligned}
\Gamma(\alpha) &= -\exp\{-x\}\, x^{\alpha-1}|_0^\infty - \int_0^\infty \left[\exp\{-x\}\,(\alpha-1)x^{\alpha-2}\, dx\ \right] \\
&= 0 + (\alpha-1)\int_0^\infty \left[\ -\exp\{-x\}\, x^{\alpha-2}\, dx\ \right] \\
&= (\alpha-1)\Gamma(\alpha-1).
\end{aligned}$$

Luego entonces, se dice que la función Gamma sigue una relación recursiva, es decir, podemos escribirla de manera más simplificada y fácil de obtener,

$$\begin{aligned}
\Gamma(\alpha) &= (\alpha-1)\Gamma(\alpha-1) \\
&= (\alpha-1)(\alpha-2)\Gamma(\alpha-2) \\
&= (\alpha-1)(\alpha-2)...\Gamma(1) \\
&= (\alpha-1)!.
\end{aligned}$$

Ahora, se define la distribución Gamma como a continuación se presenta.

Definición 1.7.10 *Sea* T *una variable aleatoria continua, cuyo recorrido son los números reales positivos. Se dice que* T *sigue un modelo de probabilidad Gamma si su función de densidad de probabilidad es de la forma,*

$$f(t) = \frac{1}{\Gamma(\alpha)}\ \beta^{\ \alpha}\ t^{\ \alpha-1}\ \exp\{-\beta t\}, \qquad con\ \ \alpha,\ \beta > 0. \tag{1.25}$$

Donde α *es el parámetro de forma,* β *el parámetro de escala y* $\Gamma(\alpha)$ *la función Gamma.*

Observación 1.7.11 *i) Si consideramos una variable aleatoria* T *que sigue un modelo Gamma con parámetros* (α, β)*, cuando* $\alpha = 1$*, entonces la variable* T *se transforma a una variable que sigue un modelo exponencial. En otras palabras tenemos que,*

$$\begin{aligned}
f(t) &= \frac{1}{\Gamma(\alpha)}\beta^\alpha t^{\alpha-1}\exp\{-\beta t\} \\
&= \beta\ \exp\{-\beta t\}.
\end{aligned} \tag{1.26}$$

Donde esta última expresión es claramente la función de densidad de una variable aleatoria que sigue un modelo exponencial con parámetro β. Así diremos entonces que la distribución exponencial es un caso especial de la distribución Gama.

ii) Esta distribución es reproductiva bajo la suma, en otras palabras, si $T_1, T_2, T_3, ..., T_n$ son variables aleatorias independientes que se distribuyen $G(\alpha, \beta)$, tenemos entonces que,

$$T_1 + T_2 + T_3 + ... + T_n \sim G(\alpha, \beta_1 + \beta_2 + \beta_3 + ... + \beta_n). \tag{1.27}$$

1.8. ¿Cómo comparar las distribuciones para elegir la mejor?

En la sección anterior se describieron algunas distribuciones que son muy útiles dentro del análisis de supervivencia, sin embargo, no son las únicas, debido a que en la teoría existe una gran variedad, cada una con características diferentes pero con un mismo enfoque, describir el comportamiento de un conjunto de datos. Resultado de la variedad que se tiene, para ajustar un modelo a un conjunto de datos es conveniente revisar que es lo que esperamos al relacionar el modelo con los datos y cuales son las ventajas que nos ofrece el modelo. Por mencionar algunos puntos básicos que se deben comparar entre distribuciones están los siguientes [1].

- Su conveniencia técnica para la inferencia estadística.
- La disponibilidad de formas explícitas y razonablemente simples para la función de supervivencia, función de densidad de probabilidad y función de riesgo.
- La capacidad de representar tanto la dispersión excesiva como la insuficiente.
- El comportamiento de la función de supervivencia en intervalos de tiempo pequeños.
- Cualquier conexión con un modelo estocástico especial de falla.

1.9. Datos censurados

Como se ha mencionado en la introducción de este capítulo, el objetivo principal del análisis de supervivencia es incorporar todos los datos que se tienen para deducir o describir como se comporta una población o grupos de individuos, sin embargo, no siempre se tienen

a disposición todos los datos que como investigador desearía tener, de modo que uno busca hallar solución al problema que surge cuando los datos se encuentran incompletos.

Una solución rápida es trabajar solamente con los datos que se tienen, sin embargo, los resultados tendrían un sesgo grande, por ello, en muchos de los casos se trabaja o se estudia los tiempos de vida o de falla censurando la información.

Denominamos datos censurados a aquellas observaciones que son incompletas, por ejemplo, si consideramos como población a un conjunto de datos de tiempo de falla, la censura existe cuando haya tiempos no definidos.

Cuando este tipo de problemas surgen en la investigación, su presunción puede ser variada, sin embargo, las causas más comunes son las siguientes:

- El objeto de estudio no presenta el evento puntual de interés dentro del periodo de estudio establecido.
- Se pierde el seguimiento del objeto durante el periodo de estudio.

1.9.1. Tipos de censura

Derivado a que este problema surge en la realidad, se han clasificado los distintos tipos de censura [1], los más comunes son,

Censura del tipo I. *El investigador recurre a este tipo de censura cuando tiene un tiempo máximo de observación, entonces aquellos elementos que no presentan la falla dentro de este periodo presentarán censura.*

Por ejemplo, consideremos a un grupo de personas con ataque de ansiedad, donde a cada uno se le aplicó un medicamento para calmarse, el tiempo de falla será aquel que comprenda el tiempo desde que se le aplicó el medicamento hasta que se calma. Además, se cuenta con un tiempo máximo de observación, digamos C, entonces si el paciente no se calma en un tiempo menor a C, este será censurado y su tiempo de falla será el tamaño máximo de observación, en este caso C. De esta manera se puede definir el tiempo de falla para cada elemento de la siguiente manera [1],

$$T = mín(T, C). \tag{1.28}$$

Censura del tipo II. *Este tipo de censura se presenta cuando la observación se extiende hasta que ocurran un cierto número, digamos K, entonces si se considera una población de n elementos, la diferencia de n y K serán datos censurados.*

Por ejemplo, si consideramos la misma población del ejemplo anterior, pero en este caso se dice que la observación de los pacientes durará hasta que al menos K hayan calmado su ataque de ansiedad, entonces al final del estudio se tendrán K tiempos de falla exactos y $n - K$ tiempos serán tiempos censurados. Así, entonces la censura está controlada por el investigador.

Censura del tipo III. *Se define a la censura del tipo III como aquellos datos que son desconocidos por alguna situación aleatoria, por lo que se podría decir que este tipo de censura es aleatoria pues a diferencia de las dos anteriores ésta no se encuentra controlada por el investigador.*

De esta manera, retomando el mismo ejemplo de los tipos de censura anteriores, los datos censurados podrían ser aquellos datos que dentro del periodo de observación fueron dejados de observar porque tal vez el individuo se fue, o en el peor de los casos murió.

Capítulo 2

Modelo de Cox

En el capítulo anterior se abordaron algunas de las distribuciones más usuales en el análisis de supervivencia, sin embargo, por ser modelos completamente paramétricos funcionan sólo para un análisis que involucran inferencias básicas, es decir, incluyen una sola distribución, pero ¿qué pasa cuando se desea hacer un análisis más real? Para estos casos, en la teoría del análisis de supervivencia se tienen diferentes modelos semi paramétricos que permiten llevar a cabo este tipo de análisis.

Entonces, tomando en cuenta que el objetivo de esta tesis es hacer un análisis completo de un conjunto de datos de tiempo de falla, el enfoque de este capítulo se centra en la definición y formulación de un modelo que por su esencia permita considerar el efecto de distintas variables explicativas. El modelo en cuestión, será el modelo de Cox, también conocido como el modelo de Riesgos Proporcionales, el cual es utilizado ampliamente en el área de supervivencia y famoso porque a diferencia de un análisis básico, este se caracteriza por considerar diferentes riesgos a los cuales se expone cualquier elemento de la población, ya que, busca representar el efecto que genera considerar un conjunto de variables explicativas en el tiempo de falla.

Entonces, con el propósito de poder comprender el modelo de Cox, se iniciará por introducir el concepto de variable explicativa, junto con su clasificación más sencilla. Seguido de esto, se planteará al modelo de Cox en su forma más básica, en el cual se consideran variables independientes del tiempo junto con parámetros que son desconocidos (estos parámetros representan el efecto de cada covariable en la tasa de riesgo).

Así, en consecuencia, a que los parámetros de las variables explicativas son desconocidos y este tipo de análisis pertenece al área de la estadística inferencial, se aborda también la descripción breve del método de máxima verosimilitud, el cual es simple y fácil de aplicar gracias a la función sencilla con la que trabaja.

2.1. Factores de riesgo

Como se mencionó en el capítulo uno, la determinación del tiempo de falla es de suma importancia, pues gracias a eso se pueden considerar diferentes escenarios a la hora de tomar decisiones. Si se considera un análisis básico el efecto de cada covariable es cero, sin embargo, el considerarlos genera un tiempo de falla más preciso, pues el cálculo de la falla se apega más a la realidad.

Al considerar variables explicativas dentro del análisis de supervivencia, se supone que para cada individuo se encuentra definido un vector, digamos Z, el cual está compuesto de k variables explicativas, es decir, $Z = (z_1, z_2, \ldots, z_k)$.

Definición 2.1.1 *Sea z_i una característica de riesgo o de vida de un individuo durante el periodo de estudio, entonces se entiende como variables explicativas a todos aquellos factores de riesgos o de vida que en algún momento afectan al tiempo de falla del elemento.*

Por ejemplo, si consideremos una población de niños entonces las variables explicativas que se pueden considerar son la edad, el sexo, su estado de salud, su situación económica, etc.

Observación 2.1.2 *Cuando se tiene un análisis sin covariables, es decir, $Z = 0$ se dice que el análisis corresponde a un conjunto de condiciones estándar [1].*

Nota: *El conjunto de variables explicativas que están relacionadas con el tiempo de falla de ahora en adelante serán denominadas "covariables".*

2.1.1. Clasificación de covariables

Las covariables se pueden clasificar en tres clases [1],

i) *Tratamientos: en esta clase se concentran aquellas variables que son de naturaleza periódica, por ejemplo, que el elemento se encuentre tomando algún tratamiento o bien que tenga un control de salud.*

ii) *Propiedades intrínsecas: aquí se encuentran las variables que son del tipo médico, demográficas e históricas. Por lo regular son aquellas características que el individuo presenta antes de ingresar al estudio como su edad, su sexo, su estado de salud, etc.*

iii) *Variables exógenas: estas variables están definidas por las características ambientales en las que los individuos se desarrollan o viven. Este tipo de covariables son excelentes para explicar el comportamiento de individuos que pertenecen a ciertos grupos que utilizan algún aparato, etc.*

Definición 2.1.3 *Sea $\psi(Z)$ la función que relaciona al vector de covariables con la supervivencia de un elemento. Definimos a β como el vector de k parámetros caracterizados de $\psi(Z)$, [1].*

Cuando en el análisis de supervivencia se tiene $\psi(Z) \neq 1$, se tienen las siguientes observaciones [1]:

1) *Un incremento en $\psi(Z)$ aumenta el riesgo al que está expuesto el elemento, por lo que su tiempo de falla es menor, es decir, su supervivencia disminuye.*

2) *Una disminución en $\psi(Z)$ disminuye el riesgo de falla del elemento, así que su tiempo de supervivencia aumenta.*

2.2. Modelo de Cox

En el área de supervivencia existen varios modelos que permiten conocer la relación que existe entre la tasa de falla de un individuo con respecto a un conjunto de variables que son medidas dentro del periodo de observación por ejemplo el modelo de Vida Acelerada, el modelo de Cox o bien el modelo de Origen Transferido, sin embargo, por ahora nos enfocaremos a plantear el modelo de Cox, también conocido como el modelo de Riesgos Proporcionales, este modelo fue planteado en el año 1985 por el Doctor David Cox, el cual permite hacer un análisis de supervivencia evaluando el efecto que genera la intervención de varias circunstancias o características que definen al objeto de estudio.

El modelo de Cox, en su formulación más simple y básica considera un vector de covariables constante, es decir, cada una de sus entradas permanece igual durante un intervalo de tiempo, pues la dependencia al tiempo es nula. De esta manera podemos enunciar la siguiente definición.

Definición 2.2.1 *Sea Z un vector de covariables constante, el modelo de riesgos proporcionales supone que la función de riesgo está definida por,*

$$h(t; Z) = \psi(Z)h_0(t), \tag{2.1}$$

donde $h_0(t)$ es la tasa de riesgo basal de cualquier elemento y $\psi(Z)$ es la función que relaciona a la función de riesgo con el vector de covariables que anteriormente ya fue definido [1].

Partiendo de la Definición 2.1.3, es claro notar que los parámetros que integran al vector β son necesarios para alcanzar el objetivo de este modelo por lo que más adelante se buscará estimarlos, entonces un candidato natural es [1],

$$\psi(Z;\beta) = \exp\left\{\beta' Z\right\}. \tag{2.2}$$

Además, si se retoma el hecho de que el conjunto de covariables definidas para cada individuo es independiente del tiempo, entonces la función planteada por el modelo se puede expresar de la siguiente manera,

$$h(t;Z) = \exp\left\{\sum_{i=0}^{k} \beta_i Z_i\right\} h_0(t). \tag{2.3}$$

Observación 2.2.2 *La función $\psi(Z) \geq 0$ siempre que el vector Z sea distinto de cero y en el caso particular donde $Z = 0$, es decir, no existan variables explicativas, la función será $\psi(Z) = \psi(0) = 1$.*

2.2.1. Ventajas del modelo de Cox

i) *Este modelo tiene una interpretación sencilla, ya que el efecto que tienen las covariables sobre los datos de tiempos de falla es multiplicar la tasa de riesgo basal por un factor constante.*

ii) *Existe soporte empírico para la suposición de riesgos proporcionales en diferentes grupos de estudio.*

iii) *La formulación del modelo se puede adecuar a datos de tiempos de falla que han sido censurados o que son únicos [1].*

Ahora que ya se ha definido el modelo de Cox, es importante mencionar que existe una condición para que este modelo sea aplicado a un conjunto de datos de tiempo de falla, sin embargo, antes es importante definir el siguiente concepto para entender dicha condición.

Definición 2.2.3 *Los riesgos definidos en el vector de covariables Z y la tasa de incidencia del grupo o individuo no expuesto a Z [1]. De esta manera, HR se puede expresar como,*

$$HR = \frac{h(t;Z)}{h(t)}. \tag{2.4}$$

La cual, puede ser entendida como una medida de efecto entre un grupo ajustado a los riesgos Z contra un grupo con riesgos nulos.

Así, de la Definición 2.2.3 se sigue que el conjunto de datos de tiempos de falla tiene que cumplir que la tasa de riesgos para cada elemento de la muestra es proporcional, es decir, la tasa de riesgos es constante en cualquier punto del tiempo.

Proposición 2.2.4 *Si se tiene una muestra de datos de tiempos de falla que se modelan con el modelo de Cox, entonces la tasa de riesgos es proporcional.*

Demostración:

Sea $T_1, T_2, ..., T_p$ una muestra de datos de tiempos de falla modelados mediante el modelo de Cox. Entonces existe un vector Z_p de covariables para cada elemento de la muestra. Luego utilizando la Definición 2.1.3, se tiene que,

$$h(t; Z) = \psi(Z_p|\beta)h_0(t).$$

Es decir, $h(t|Z_p)$ es la función de riesgo para el elemento si se consideran las Z_p covariables. Entonces, si $Z_p = 0$ tenemos que,

$$h(t|Z_p) = h_0(t).$$

Luego entonces,

$$\begin{aligned} HR &= \frac{h(t|Z_p)}{h(t|Z_p = 0)} \\ &= \frac{\psi(Z_p)h_0(t)}{h_0(t)} \\ &= \psi(Z_p|\beta). \end{aligned}$$

Con lo cual queda demostrado, debido a que $\psi(Z_p|\beta)$ no depende del tiempo. ■

Finalmente, ya planteado el modelo de Cox, lo siguiente es conocer el método de estimación que se utilizará para estimar los parámetros del vector β.

Dentro del área de la estadística inferencial se pueden hallar varios métodos de estimación, pero en esta ocasión se utilizará el método de máxima verosimilitud parcial, el cual nos lleva primero a definir al método de máxima verosimilitud clásico.

2.3. Método de máxima verosimilitud

El método de máxima verosimilitud es un método de estimación que utiliza la función de densidad de probabilidad conjunta, además supone que se tiene una muestra donde cada elemento es independiente e idénticamente distribuido.

Definición 2.3.1 *Sea $T_1, T_2, \ldots, T_n$ los tiempos de falla de una muestra de n variables aleatorias independientes e idénticamente distribuidas. La función de máxima verosimilitud se define como la función de densidad conjunta [5], esto es,*

$$f(t_1, t_2, ..., t_n; \psi) = f(t_1; \phi) f(t_2; \psi) ... f(t_n; \psi). \tag{2.5}$$

Nota: *A la función anterior se le denota por $L(\psi; t_1, t_2, ..., t_n)$, donde $\psi = (\theta_1, \theta_2, ..., \theta_k)$ representa a los k parámetros que se desean estimar.*

De manera que la función de máxima verosimilitud se interpreta como la probabilidad de que los tiempos de falla $T_1, T_2, \ldots, T_n$ tomen los valores $t_1, t_2, \ldots, t_n$ dado los parámetros ψ esto es, que cada tiempo de falla contribuye con la probabilidad $f_i(t; \psi)$ a la densidad del tiempo de falla en t.

El método de máxima verosimilitud consiste en tomar a $L(\psi; t_1, t_2, ..., t_n)$ y determinar los parámetros que maximizan dicha función, es decir, se busca hallar los parámetros $\tilde{\psi}$ que al ser sustituidos en $L(\psi; t_1, t_2, ..., t_n)$ nos den la máxima probabilidad de que los tiempos de falla $T_1, T_2, \ldots, T_n$ tomen los valores $t_1, t_2, \ldots, t_n$.

Algoritmo 2.3.2 *Para describir el procedimiento a seguir, supóngase que se tiene una muestra de n tiempos de falla, digamos $T_1, T_2, \ldots, T_n$ todos independientes e idénticamente distribuidos, además supóngase que no existe censura de datos. Entonces empleando la expresión que define a la función de máxima verosimilitud, se tiene que*

$$\begin{aligned} L(\psi; t_1, t_2, ..., t_n) &= f(t_1; \psi) f(t_2; \psi) f(t_n; \psi) \\ &= \prod_{i=1}^{n} f(t_i; \psi). \end{aligned} \tag{2.6}$$

Luego por la simplicidad de trabajar con logaritmos, tenemos que,

$$\begin{aligned} l(\psi) = l(\theta_1, \theta_2, ..., \theta_k) &= \ln \left[L(\psi; t_1, t_2, ..., t_n) \right] \\ &= \sum_{i=1}^{n} \ln \left[f(t_i; \psi) \right]. \end{aligned} \tag{2.7}$$

Así, los estimadores de máxima verosimilitud, serán las soluciones a las k ecuaciones siguientes [5],

$$\frac{\partial l(\theta_1, \theta_2, ..., \theta_k)}{\partial \theta_1} = 0,$$

$$\frac{\partial l(\theta_1, \theta_2, ..., \theta_k)}{\partial \theta_2} = 0,$$

$$\vdots$$

$$\frac{\partial l(\theta_1, \theta_2, ..., \theta_k)}{\partial \theta_k} = 0.$$

Por otro lado, si dejamos de considerar que todos los tiempos de falla son únicos, entonces la función de máxima verosimilitud al considerar censura queda determinada de la siguiente manera [1],

$$L(\psi) = \prod_u f(t_i; \psi) \prod_c S(C_i; \psi).$$

$$L(\psi) = \prod_u f(x_i; \psi) \prod_c S(x_i; \psi).$$

Luego tenemos que,

$$l(\psi) = \sum_u \ln\left[f(x_i; \psi)\right] + \sum_c \ln\left[S(x_i; \psi)\right].$$

Puesto que $f(t) = h(t)S(t)$, lo anterior es igual a,

$$\begin{aligned} l(\psi) &= \sum_u \ln\left[h(x_i; \psi)S(x_i; \psi)\right] + \sum_c \ln\left[S(x_i; \psi)\right] \\ &= \sum_u \ln\left[h(x_i; \psi)\right] + \sum_u \ln\left[S(x_i; \psi)\right] + \sum_c \ln\left[S(x_i; \psi)\right] \\ &= \sum_u \ln\left[h(x_i; \psi)\right] + \sum_n \ln\left[S(x_i; \psi)\right]. \end{aligned} \tag{2.8}$$

Una vez obtenida la igualdad anterior, los estimadores se obtienen de la misma manera como en el caso donde no se considera censura.

El método de máxima verosimilitud descrito anteriormente se contempla para aquellos modelos paramétricos, en el caso del modelo de Cox el cuál es un modelo semiparamétrico no es adecuado, pues como se definió al modelo anteriormente este solo tiene una parte paramétrica en consecuencia, se describe el método de máxima verosimilitud con una pequeña adecuación, al cual se le denomina método de máxima verosimilitud parcial, y cuyo planteamiento se describe a continuación.

2.4. Método de máxima verosimilitud parcial

Antes de enunciar, al método de máxima verosimilitud parcial para el modelo de Cox es relevante remarcar que el objetivo de este análisis se centra en hacer inferencia sobre el vector de parámetros β. Entonces, para iniciar se considera el caso en el cual se supone que no existe censura en el conjunto de datos de tiempo de falla.

Entonces para obtener la función de verosimilitud considérese una muestra de n elementos, en este caso n tiempos de falla, cada uno único. Entonces, sean $t_1 < t_2 < ... < t_n$ los tiempos de falla de los n individuos donde $t_i = t_j$ si y sólo si $i = j$, lo anterior para asegurar que dentro de los datos no existen empates. También definamos a γ_i la persona que falla en el tiempo t_i. Antes de continuar se define lo siguiente.

Definición 2.4.1 *Sea $R(t_i)$ el conjunto de todos los elementos de la muestra que no han fallado hasta el tiempo t_i, es decir, su tiempo de falla es mayor que t_i. Así, $R(t_i)$ se puede expresar como [1],*

$$R(t_i) = \{\gamma_j | t_j > t_i\}. \tag{2.9}$$

Recordemos que por el momento queremos obtener la función de verosimilitud, para ello necesitamos la distribución de probabilidad conjunta, en este caso buscamos a $P[\, T_1, T_2, ..., T_n\,]$, la cual se puede obtener realizando todas las permutaciones posibles en $(1, \ldots, n)$ dado que los tiempos de falla que se tiene están ordenados, pero si consideramos que cada tiempo de falla en la muestra es idénticamente distribuido e independiente, lo que en realidad buscamos es $P[\, \gamma_j | \gamma_1, \gamma_2, ..., \gamma_(j-1)\,]$, es decir, buscamos la probabilidad de que γ_j falle en t_j dado que un individuo del conjunto de riesgos $R(t_i)$ falló en t_i,en otras palabras lo que buscamos es,

$$P[\, \gamma_j | \gamma_1, \gamma_2, ..., \gamma_{j-1}\,] = \frac{\psi(t_j)}{\sum\limits_{k \in R(t_j)} \psi(k)}. \tag{2.10}$$

Pero entonces la función de distribución conjunta por definición queda expresada como,

$$\begin{aligned} P\,[\ T_1, T_2, ..., T_n\ ] &= \prod_{i=1}^{n} P[\ \gamma_j | \gamma_1, \gamma_2, ..., \gamma_{j-1}\ ] \\ &= \prod_{j=1}^{n} \frac{\psi(t_j)}{\sum\limits_{k \in R(t_j)} \psi(k)}. \end{aligned} \tag{2.11}$$

Una vez obtenida la función de verosimilitud, lo que sigue es obtener la función de verosimilitud considerando censura en los tiempos de falla, la cual es obtenida de manera similar si se supone que los datos censurados se encuentran a la derecha de todos los tiempos de falla exactos.

Así, si en la muestra de elementos consideramos que existen algunos tiempos que han sido censurados, digamos k tiempos son exactos y el resto $n - k$ son datos censurados, entonces tenemos $t_1, t_2, ..., t_d$ tiempos de falla ordenados y sin empates. Además, nuevamente definamos a γ_j cómo el sujeto que falla en el tiempo t_j donde $j \in (1, ..., d)$ y sea $R(t_j)$ el correspondiente conjunto de riesgos, donde este conjunto sólo incluye los tiempos de falla que ocurren en el intervalo de tiempo $(0, t_j)$. Por lo que la función de verosimilitud es,

$$L(T_1, T_2, ..., T_d; \beta) = \prod_{i=1}^{d} \frac{\psi(t_j)}{\sum\limits_{k \in R(t_j)} \psi(k)}. \tag{2.12}$$

De modo que en la función de verosimilitud los datos que son censurados han sido omitidos, pues en el conjunto de riesgos considerado anteriormente omite la participación de los mismos en cada conjunto. Pero entonces el mecanismo de censura no depende de β, lo que nos lleva a decir que para la inferencia de β también pueden ser omitidos.

Obtenida la función de verosimilitud parcial, lo que continúa es derivar la misma para obtener la inferencia de los parámetros de las covariables. Aquí es indispensable que dentro de los datos existan tasas proporcionales y además, que para todo $t_i \in \phi(t_i)$ tenga definida su primera y segunda derivada, lo que es equivalente a decir que,

$$\frac{\partial \psi(T_i)}{\partial \beta_r} = \psi_r(T_i)$$

$$\frac{\partial^2 \psi(T_i)}{\partial \beta_r \partial \beta_s} = \psi_{rs}(T_i).$$

Luego, como al aplicar logaritmos en el método de máxima verosimilitud permite obtener una expresión más simple, en este modelo también se obtienen los logaritmos, por lo que la

función queda de la siguiente manera,

$$\begin{aligned} l &= \ln [\ L(T_1, T_2, ..., T_d; \beta)\ ] \\ &= \ln \left[\prod_{j=1}^{d} \frac{\psi(t_j)}{\sum\limits_{k \in R(t_j)} \psi(k)} \right] \\ &= \sum_{j=1}^{d} \left\{ \ln [\ \psi(t_j)\ ] - \ln \left[\sum_{k \in R(t_j)} \psi(k) \right] \right\}. \end{aligned} \tag{2.13}$$

Pero entonces la primera y segunda derivada quedan definidas de la siguiente manera,

$$\begin{aligned} \frac{\partial l}{\partial \beta_r} &= \frac{\partial}{\partial \beta_r} \left\{ \sum_{i=1}^{d} \left(\ln [\psi(T_i)] - \ln \left[\sum_{k \in R(T_i)} \psi(k) \right] \right) \right\} \\ &= \sum_{i=1}^{d} \left(\frac{\psi_r(T_i)}{\psi(T_i)} - \frac{\sum\limits_{k \in R(T_i)} \psi_r(k)}{\sum\limits_{k \in R(T_i)} \psi(k)} \right). \end{aligned} \tag{2.14}$$

Así los estimadores del vector de parámetros β se obtienen resolviendo las ecuaciones simultáneas de

$$\frac{\partial l}{\partial \beta_r},$$

como se hace en el método de máxima verosimilitud clásico.

En la definición del modelo de Cox, se mencionó que para la función $\psi(Z|\beta)$ existen diferentes expresiones con las cuales ésta puede ser remplazada, pero si tomamos al más natural de los candidatos, el remplazo sería una función lineal como la siguiente,

$$\psi(Z|\beta) = \exp \left\{ \sum_{i=1}^{p} \beta' Z \right\}.$$

La cual, si es remplazada en las expresiones que se obtuvieron al derivar a la función de máxima verosimilitud parcial, estas se simplifican a una manera más simple.

Capítulo 3

Modelo de Cox con un punto de cambio

En el análisis de supervivencia es usual encontrarse con complicaciones al ajustar un conjunto de datos de tiempo de falla, por ejemplo, que algunas observaciones sean vulnerables a sufrir algún cambio dentro del periodo de estudio, de manera que el modelo elegido no podría ser la mejor opción para ajustar el conjunto de datos, es decir, el investigador podría pensar que el modelo seleccionado no es el adecuado.

Este tipo de complicaciones están presentes principalmente en el área de las ciencias de la salud, como lo es la medicina, biología, química, etc. Ya que en estos campos uno de los objetivos convencionales es estudiar un mismo experimento bajo condiciones diferentes o bajo las mismas condiciones. Por ello surge la necesidad de considerar la existencia de un punto de cambio en la función de riesgo, debido a que el rol de esta función en el análisis de supervivencia es uno de los más importantes.

En este capítulo se propone el modelo de Cox con un punto de cambio, el cual considera un incremento o decremento en la i-ésima covariable, además se obtendrá la función de supervivencia y la función de densidad de probabilidades correspondientes a este modelo.

Cuando se estudia a un conjunto de datos de tiempos de falla que satisfagan la existencia de punto de cambio en la distribución no es necesario tener un modelo de supervivencia básica, debido a que el investigador ajusta cualquier modelo bajo esta condición, así que, como el objetivo de esta tesis es ajustar a un conjunto de datos al modelo de Cox con punto de cambio, el cual se definirá más adelante, se iniciará planteando un modelo paramétrico que considere un punto de cambio con la finalidad de entender con una mejor precisión esta consideración.

Cuando se considera la existencia de un punto de cambio en consecuencia se tiene que la función de riesgo deja de ser única para todos los elementos, es decir, si se estudia una muestra de tamaño n, de tiempos de falla digamos $T_1, T_2, \ldots, T_n$, cierta parte de esta muestra estarán asociados a una tasa de riesgo, digamos k, mientras que el resto estarán sujetos a otra. Así damos paso al modelo exponencial con punto de cambio.

3.1. Modelo exponencial con un punto de cambio

Dentro de los estudios de supervivencia, cuando se opta por ajustar al conjunto de datos bajo un modelo paramétrico el investigador se expone a que la naturaleza de la muestra no cumpla con las condiciones necesarias del modelo que se ha seleccionado, sin embargo, por conveniencia el investigador pudiera considerar un modelo paramétrico que considere un punto de cambio.

El modelo paramétrico con punto de cambio generalmente usado en el análisis de conjuntos de datos con un punto de cambio es el modelo Exponencial, ya que es bien sabido que este modelo, como tal es uno de los modelos más importantes en el área del análisis de supervivencia, por que su tasa de riesgo constante e independiente del tiempo facilita cualquier cálculo.

Para definir la función de riesgo asociada a este modelo, considérese a T como una variable aleatoria que sigue una distribución exponencial con punto de cambio, entonces se tiene la siguiente definición.

Definición 3.1.1 *Sea T una variable aleatoria continua que representa un tiempo de falla y que además sigue una distribución exponencial con punto de cambio. La función de riesgo asociada a T, digamos $\lambda(t)$ está dada por [4] ,*

$$\lambda(t) = \begin{cases} \lambda_1, & \text{si } t \leq \tau, \\ \lambda_2, & \text{si } t > \tau. \end{cases} \tag{3.1}$$

Donde λ_1 y λ_2 representan las tasas de riesgo en $[\,0, \tau\,]$ y (τ, ∞) respectivamente, y además τ representa el punto de cambio en la función de riesgo, el cual es desconocido.

De la definición de la función de riesgo asociada a T es fácil obtener la función de supervivencia y la función de densidad de probabilidad, haciendo uso de algunos resultados del Capítulo 1, como se muestra en los siguientes teoremas.

Teorema 3.1.2 *Sea T una variable aleatoria continua con distribución exponencial con un punto de cambio, entonces la función de supervivencia asociada a T es,*

$$S(t) = \begin{cases} \exp\{-\lambda_1 t\}, & \text{para } t \leq \tau, \\ \exp\{-\lambda_1 t - \lambda_2(t-\tau)\}, & \text{para } t > \tau. \end{cases} \tag{3.2}$$

Demostración:

Del Teorema 1.5.3 se tiene que,

$$S(t) = \exp - \left\{ \int_0^t \lambda(x)\, dx \right\}.$$

Entonces si $t \leq \tau$, se tiene que

$$\begin{aligned} S(t) &= \exp\left\{ - \int_0^t \lambda_1\, dx \right\} \\ &= \exp\{-\lambda_1 t\}. \end{aligned}$$

Luego si $t > \tau$, la función de supervivencia es,

$$\begin{aligned} S(t) &= \exp\left\{ - \int_0^\tau \lambda_1\, dx - \int_\tau^t \lambda_2\, dx \right\} \\ &= \exp\{-\lambda_1 \tau - \lambda_2(\tau - t)\}. \end{aligned}$$

Por lo tanto, se llega a que,

$$S(t) = \begin{cases} \exp\{-\lambda_1 t\}, & \text{para } t \leq \tau, \\ \exp\{-\lambda_1 t - \lambda_2(t-\tau)\}, & \text{para } t > \tau. \end{cases}$$

Con lo cual queda demostrado. ■

Teorema 3.1.3 *Sea T una variable aleatoria que sigue una distribución exponencial con un punto de cambio, entonces su función de densidad de probabilidad es,*

$$f(t) = \begin{cases} \lambda_1 \exp\{-\lambda_1 t\}, & \text{para } t \leq \tau, \\ \lambda_2 \exp\{-\lambda_1 t - \lambda_2(t-\tau)\}, & \text{para } t > \tau. \end{cases} \tag{3.3}$$

Demostración:

Del Lema 1.5.2 tenemos que,

$$f(t) = \lambda(t)S(t).$$

Así que si $t \leq \tau$, se tiene que la función de densidad de probabilidad es,

$$f(t) = \lambda_1 \exp\{-\lambda_1 t\}.$$

Mientras que si $t > \tau$, se tiene que

$$f(t) = \lambda_1 \exp\{-\lambda_1 t - \lambda_2(t-\tau)\}.$$

Por lo tanto,

$$f(t) = \begin{cases} \lambda_1 \exp\{-\lambda_1 t\}, & \text{para } t \leq \tau, \\ \lambda_2 \exp\{-\lambda_1 t - \lambda_2(t-\tau)\}, & \text{para } t > \tau. \end{cases} \blacksquare$$

3.2. Modelo de Cox con un punto de cambio

Cuando el estudio de un conjunto de datos de tiempos de falla se encuentra vulnerable a cambios en su distribución y además no se sabe con exactitud la naturaleza del conjunto, es difícil ajustar dicho conjunto de datos a un modelo de supervivencia paramétrico. Por ello la propuesta es plantear el modelo de Cox considerando un punto de cambio, pues como se definió en el Capítulo 2 este modelo, es un modelo semiparamétrico que permite llevar a cabo la estimación del efecto de las covariables sin tener especificada la función de riesgo basal que está asociada a la función de riesgo de cada elemento de la muestra.

Para considerar un punto de cambio en la función de riesgo de este modelo, suponemos que del vector de covariables $Z_k = (z_1, z_2, \ldots, z_k)$ asociado a cada tiempo de falla T, se tiene un incremento o decremento en una única covariable, digamos z_i, mientras que el resto se mantiene igual durante todo el periodo de tiempo.

Recuérdese que de la Definición 2.2.1 se tiene que la función de riesgo para el modelo de Cox tradicional es,

$$h(t|Z) = h_0(t) \exp\{\beta_k Z_k\}.$$

Entonces para plantear a este modelo con un punto de cambio, se tiene la siguiente definición.

Definición 3.2.1 *Sea T una variable aleatoria que representa un tiempo de falla y además sigue el modelo de Cox con un punto de cambio, es decir, también existe un vector de covariables Z_k. Se define a la función de riesgo asociada a T, digamos $h(t|Z)$ como,*

$$h(t|Z) = \begin{cases} h_0(t) \exp\left\{\beta_k' Z_k\right\}, & \text{para } t \leq \tau, \\ h_0(t) \exp\left\{(\beta_k + \theta_k)' Z_k\right\}, & \text{para } t > \tau. \end{cases} \tag{3.4}$$

Donde θ_k representa el cambio en la i-ésima covariable, es decir, la única entrada distinta de cero será la i-ésima, τ representa el punto de cambio en la función de riesgo, y como se ha mencionado anteriormente $h_0(t)$ representa la función de riesgo basal.

Después de definir a la función de riesgo para el modelo propuesto, lo que nos interesa ahora es conocer la función de supervivencia así como la función de densidad de probabilidad que más adelante serán necesarias para hacer las estimaciones correspondientes a los parámetros desconocidos.

Nota: *Con el fin de facilitar los cálculos, se define a λ_0 como la función de riesgo basal para el modelo de Cox, cuyo valor ya es conocido.*

Teorema 3.2.2 *Sea T una variable aleatoria que representa un tiempo de vida o falla, y que además tiene una función de riesgo como se ha dado en la Definición 3.2.1. Entonces la función de supervivencia asociada a T es,*

$$S(t|Z) = \begin{cases} \exp\left\{-t\lambda_0 \exp\left(\beta_k' Z_k\right)\right\}, & \text{para } t \leq \tau, \\ \exp\left\{-\left[\tau\lambda_0 \exp\left(\beta_k' Z_k\right)\right] - \left[(t-\tau)\lambda_0 \exp\left((\beta_k + \theta_k)' Z_k\right)\right]\right\}, & \text{para } t > \tau. \end{cases} \tag{3.5}$$

Demostración:

Para está demostración se utilizan argumentos análogos a los utilizados en el Teorema 3.1.2, por lo que del Resultado 1.5.3 tenemos que,

$$S(t) = \exp - \left\{\int_0^t \lambda(x)\, dx\right\}.$$

Entonces si $t \leq \tau$, se tiene que

$$\begin{aligned} S(t|Z) &= \exp\left\{-\int_0^t h_0(t) \exp\left(\beta_k' Z_k\right)\, dx\right\} \\ &= \exp\left\{-\int_0^t \lambda(t) \exp\left(\beta_k' Z_k\right)\, dx\right\} \\ &= \exp\left\{-t\lambda_0 \exp\left(\beta_k' Z_k\right)\right\}. \end{aligned}$$

Luego si $t > \tau$, la función de supervivencia es,

$$\begin{aligned} S(t|Z) &= \exp\left\{-\int_0^\tau h_0(t)\exp\left(\beta_k' Z_k\right)\,dx - \int_\tau^t h_0(t)\exp\left\{(\beta_k+\theta_k)'Z_k\right\}\,dx\right\} \\ &= \exp\left\{-\int_0^\tau \lambda_0\exp\left(\beta_k' Z_k\right)\,dx - \int_\tau^t \lambda_0\exp\left\{(\beta_k+\theta_k)'Z_k\right\}\,dx\right\} \\ &= \exp\left\{-\left[\tau\lambda_0\exp\left(\beta_k' Z_k\right)\right] - \left[(t-\tau)\lambda_0\exp\left((\beta_k+\theta_k)'Z_k\right)\right]\right\}. \end{aligned}$$

Por lo tanto, se llega a que,

$$S(t|Z) = \begin{cases} \exp\left\{-t\lambda_0\exp\left(\beta_k' Z_k\right)\right\}, & \text{para } t \le \tau, \\ \exp\left\{-\left[\tau\lambda_0\exp\left(\beta_k' Z_k\right)\right] - \left[(t-\tau)\lambda_0\exp\left((\beta_k+\theta_k)'Z_k\right)\right]\right\}, & \text{para } t > \tau. \end{cases}$$

lo cual era lo requerido. ■

Ahora para determinar la función de densidad de probabilidad asociada a T, se requiere del siguiente teorema.

Teorema 3.2.3 *Sea T una variable aleatoria que representa un tiempo de falla y que además tiene una función de riesgo como la que se ha dado en la Definición 3.2.1. Entonces la función de densidad de probabilidad asociada a T es,*

$$f(t|Z) = \begin{cases} \lambda_0\exp\left\{\beta_k' Z_k\right\}\exp\left\{-t\lambda_0\exp\left(\beta_k' Z_k\right)\right\}, & \text{si } t \le \tau, \\ \lambda_0\exp\left\{(\beta_k+\theta_k)'Z_k\right\}\exp\left\{-\left[\tau\lambda_0\exp\left(\beta_k' Z_k\right)\right]\right\} * & \\ \exp\left\{-\left[(t-\tau)\lambda_0\exp\left((\beta_k+\theta_k)'Z_k\right)\right]\right\}, & \text{si } t > \tau. \end{cases} \tag{3.6}$$

Demostración:

Del Lema 1.5.2 tenemos que,

$$f(t) = \lambda(t)S(t).$$

Así que si $t \le \tau$, se tiene que la función de densidad de probabilidad es,

$$f(t|Z) = \lambda_0\exp\left\{\beta_k' Z_k\right\}\exp\left\{-t\lambda_0\exp\left(\beta_k' Z_k\right)\right\}.$$

Mientras que si $t > \tau$, se tiene que

$$f(t|Z) = \lambda_0\exp\left\{(\beta_k+\theta_k)'Z_k\right\}\exp\left\{-\left[\tau\lambda_0\exp\left(\beta_k' Z_k\right)\right] - \left[(t-\tau)\lambda_0\exp\left((\beta_k+\theta_k)'Z_k\right)\right]\right\}.$$

Por lo tanto,

$$f(t|Z) = \begin{cases} \lambda_0 \exp\left\{\beta_k' Z_k\right\} \exp\left\{-t\lambda_0 \exp\left(\beta_k' Z_k\right)\right\}, & \text{si } t \leq \tau, \\ \lambda_0 \exp\left\{(\beta_k + \theta_k)' Z_k\right\} \exp\left\{-\left[\tau\lambda_0 \exp\left(\beta_k' Z_k\right)\right]\right\} * & \\ \exp\left\{-\left[(t-\tau)\lambda_0 \exp\left((\beta_k + \theta_k)' Z_k\right)\right]\right\}, & \text{si } t > \tau. \end{cases}$$

■

Capítulo 4

Punto de cambio a través de ensayo y error del modelo de Cox

En los capítulos anteriores se abordó el modelo de Cox, un modelo semiparamétrico que nos permite estudiar tiempos de falla hasta que ocurre un evento de interés considerando el efecto que genera un conjunto de variables asociadas al elemento de estudio, denominado anteriormente covariables. Actualmente existen diversas áreas donde la aplicación de Cox funciona como uno de los mejores modelos para describir situaciones de interés, principalmente en el área médica, sin embargo, en este capítulo se aplicará dicho modelo a una muestra de datos del ámbito financiero.

El objetivo es aplicar uno de los modelos más importantes del área de supervivencia a un conjunto de datos financieros que permita conocer cómo se comportan los datos a partir de la teoría básica, es decir, teniendo un conjunto de tiempos de falla se pueda calcular el riesgo que corre un elemento de sufrir el evento de interés sumando los efectos que las características asociadas a él puedan generar.

La muestra de datos con la que se trabajará proviene de una AFORE del mercado, donde uno de los principales problemas es la tasa de deserción dentro de los trabajadores afiliados, la aplicación consistirá en conocer cómo influyen las características de cada trabajador en el tiempo que permanecen dentro de la empresa, cuando nos referimos al tiempo, el estudio considera tiempo de falla al tiempo comprendido desde que el trabajador se suma a la empresa hasta que este es pensionado o traspasa su cuenta de ahorro individual a otra AFORE, para continuar con nuestro objetivo a continuación se describe de manera más detallada la muestra.

4.1. Descripción de la muestra de datos

Para aplicar el modelo de Cox usaremos una muestra que pertenece a una base de datos que contiene los registros de trabajadores que se encuentran afiliados a una AFORE del mercado, dentro de estos registros se pueden hallar a su vez diversas características para cada trabajador afiliado que cuyos valores fueron capturados al momento de su registro en la empresa.

La muestra de datos con la que se trabaja pertenece a la AFORE Coppel, la cual contiene 2888 observaciones, cada una corresponde a un trabajador afiliado al sistema de AFORE y para cada uno se tienen 8 variables asociadas contando el tiempo de falla, las cuales describimos a continuación.

- Tiempo
  Es una variable de tipo numérico y es la más importante en el modelo, representa los tiempos de falla, en este caso los tiempos de estancia dentro de la empresa antes de que el trabajador solicite su traspaso a una AFORE competencia.

  Puntualmente en esta variable se guarda el tiempo que transcurrió desde que el trabajador se registró en la empresa hasta el día en el que solicitó su traspaso, es decir, hasta que la empresa traspasa su cuenta de ahorro individual a otra AFORE competencia.

- Censura
  La variable censura es un flag que nos indica si el trabajador ha sido traspasado o no, en el caso negativo se dirá que ese trabajador mantiene su cuenta de ahorro individual en la empresa, la variable toma los valores 1 y 0 respectivamente.

- Género
  Variable de tipo flag que únicamente toma dos valores: 0 ó 1, el 0 nos dice que el trabajador es "Hombre" mientras que el 1 representa que el género del trabajador es "Mujer".

- Edad
  Esta variable nos dice la edad que tiene la persona al momento de afiliarse a la empresa, los valores que toma la variable dentro de la muestra se mueven en el rango (18,50) años, además, por la naturaleza esta variable es de tipo numérico - continua y no negativa.

- Alta
  Representada por un flag que tomas valores de 0 ó 1, nos permite identificar si el trabajador es un registro o un traspaso, cuando hablamos de registro nos referimos a que el trabajador se registró por primera vez en el sistema AFORE, mientras que si es un traspaso se entiende que el trabajador ya estaba registrado en alguna otra AFORE pero decidió cambiarse a la nuestra.

- Flag-Saldo
  La variable Flag Saldo se refiere al saldo para el retiro con el que cuenta el trabajador para pensionarse, en este caso tomará el valor de 0 cuando el saldo del trabajador sea cero y en caso contrario 1.

- Flag-AporP
  Esta variable hace referencia a las aportaciones patronales que ha tenido el trabajador durante los últimos 12 meses, para nuestro caso esta variable sólo será un flag que toma el valor de 0 cuando se haya mantenido inactivo dentro de un trabajo formal y 1 cuando se hayan realizado aportaciones patronales durante los últimos 12 meses.

- Flag-Retiros
  De tipo flag, nos permite identificar que trabajadores hicieron retiros durante los últimos 12 meses, es decir, cuando la variable toma el valor 0 decimos que no se hizo retiro alguno durante los últimos 12 meses y 1 en caso contrario.

A continuación se presentan algunos de los registros que podemos encontrar dentro de la muestra.

	TIEMPO	CENSURA	GENERO	EDAD	ALTA	FLAGSALDO	FLAGAPORP	FLAGRETIROS
1	11	0	2	38	0	1	1	0
2	3	0	2	47	1	1	1	0
3	7	0	1	25	0	1	1	0
4	3	0	2	23	0	1	1	0
5	5	1	1	28	0	1	1	0
6	4	0	1	26	0	1	1	0
7	7	0	1	32	0	1	1	0
8	3	0	1	21	0	1	1	0
9	4	0	2	29	1	1	1	0
10	8	0	2	38	0	1	1	1
11	6	0	2	49	1	1	1	0
12	2	0	2	34	0	1	1	0
13	3	0	1	45	1	1	1	0
14	5	0	1	30	0	1	1	0
15	3	0	1	26	0	1	1	0
16	3	0	1	51	1	1	1	0
17	4	0	1	51	1	1	1	0
18	10	0	2	42	1	1	1	1
19	3	0	1	56	0	1	1	0
20	2	0	2	23	0	1	1	0
21	9	0	1	31	0	1	1	0
22	3	0	2	41	1	1	1	0
23	3	0	1	26	1	1	1	0
24	5	0	1	26	0	1	1	0
25	3	0	2	57	1	1	1	0

Figura 4.1: Muestra de la base de datos de AFORE.

4.2. Aplicación del modelo de Cox

Antes de generar el modelo de Cox con un punto de cambio, generamos el modelo clásico de Cox, para analizar cómo se comportan los datos, además, servirá para seleccionar que variables de las que se tienen disponibles describen mejor el evento de interés. Como se mencionó en el primer capítulo antes de generar el modelo es importante definir el evento de interés, por ello para lo que concierne nuestra aplicación definimos lo siguiente:

- Origen: el punto en el que inicia nuestro estudio es cuando el trabajador se integra a la empresa, dentro del sistema de AFORE.
- Escala para medir el tiempo: la unidad de medida que usaremos es el tiempo de estancia en la empresa en años.
- Concepción del evento de falla: el evento de interés es cuando el trabajador se traspasa a otra AFORE competencia, por lo que después de este punto se considera inactivo. En el caso donde el evento de interés no haya ocurrido, es decir, la cuenta del trabajador siga bajo el control de la empresa, el tiempo será censurado.

Para generar el modelo se usará el lenguaje de programación R.

4.2.1. Estimación del modelo de Cox: Riesgos proporcionales

Para generar el modelo, en primer lugar agregaremos todas variables disponibles dentro la muestra y una vez generado se analizará cómo se comportan los datos para poder elegir que variables tienen un nivel de significancia mayor para describir nuestro evento de interés. Después se volverá a generar el modelo con las variables que mejor describen los tiempos de falla, se revisará el ajuste del modelo y se verificará si se cumple con el supuesto de riesgos proporcionales.

Así entonces, buscamos estimar lo siguiente:

$$h(t; Z_6) = \exp\left\{\sum_{i=1}^{6} \beta_i Z_i\right\} h_0(t).$$

Donde

$Z_6 = (Z_1, Z_2, Z_3, Z_4, Z_5, Z_6) = (Género, Edad, Alta, Flag{-}Saldo, Flag{-}AporP, Flag{-}Retiros)$.

Utilizando la paquetería de R, obtenemos los valores estimados para las β_i correspondientes:

Covariable	Coeficiente β	Exp(Coeficiente β)	se(Coeficiente β)	Pr(>| z |)
Género	-0.0977	0.9069	0.0949	0.3030
Edad	-0.0335	0.9671	0.0050	1.62E-10
Alta	0.6107	1.8417	0.1078	1.45E-07
Saldo	-3.2850	0.0374	0.1533	2.00E-15
Aporp12	-0.0779	0.9250	0.1990	0.5950
Retiros	-15.8100	0.0000	110.6000	0.9890

Figura 4.2: Coeficientes para el modelo de Cox considerando todas las covariables.

De la tabla que se muestra en la Figura 4.2 podemos verificar que no todas las variables son significativas, basándonos en la probabilidad de que cada covariable sea cero las únicas que describen mejor el tiempo de falla que nos interesa son la *Edad*, *Alta* y $Flag - Saldo$. De manera que se volverá a generar el modelo de Cox únicamente con estas tres covariables. Los resultados son los siguientes:

Covariable	Coeficiente β	Exp(Coeficiente β)	se(Coeficiente β)	Pr (> | z |)
Edad	-0.0633	0.9386	0.0041	2.00E-16
Alta	1.0900	2.9742	0.0751	2.00E-16
Flag- Saldo	-2.3202	0.0982	0.1331	2.00E-16

Figura 4.3: Coeficientes para el modelo de Cox con las covariables más significativas.

Covariable	Límite Inferior al .95%	Límite superior al .955
Edad	0.9312	0.9461
Alta	2.5672	3.4458
Flag- Saldo	0.0757	0.1275

Figura 4.4: Intervalos de confianza al 95 % de probabilidad.

En la tabla de la Figura 4.3 se puede visualizar el valor de las β_i correspondientes a las covariables *Edad*, *Alta* y $Flag - Saldo$, además, el valor del $p - value$ nos hace rechazar la hipótesis nula de que las covariables son cero. Así entonces, la función de riesgo de Cox se define de la siguiente manera:

$$\begin{aligned} h(t; Z_3) &= \exp\left\{\sum_{i=1}^{3} \beta_i Z_i\right\} h_0(t) \\ &= \exp\left\{\beta_1(Edad) + \beta_2(Alta) + \beta_3(Flag - Saldo)\right\} h_0(t) \\ &= \exp\left\{-0.063(Edad) + 1.090(Alta) - 2.320(Flag - Saldo)\right\} h_0(t). \end{aligned}$$

Lo resultados anteriores nos llevan a decir que el riesgo relativo de los trabajadores con respecto a su edad aumenta en 0.938 manteniendo las otras dos covariables constantes, mientras que si el trabajador es un traspaso previo el riesgo de irse es 2.974 más con respecto a los trabajadores que inician con su cuenta individual, por otro lado, si los trabajadores cuentan con un saldo distinto de cero el riesgo de ser traspasados a otra AFORE aumenta en un 0.098.

Descritos los resultados previos, revisaremos el ajuste del modelo haciendo uso del test de máxima verosimilitud, el test de Wald y la prueba Score, de donde obtenemos los siguientes resultados:

Coeficiente de Determiandción	0.8
Razón de verosimilitud	2.00E-16
Test de Walt	2.00E-16
Prueba de Score	2.00E-16

Figura 4.5: Ajuste del modelo.

En la Figura 4.5 se muestra que el coeficiente de determinación, el cual es bastante bueno, ya que al tener el 0.80 de ajuste permite asegurar que las covariables introducidas describen bastante bien el riesgo de que los trabajadores dejen la empresa, es este caso sean traspasados a otra AFORE.

El valor del $p-value$ para la prueba de máxima verosimilitud nos indica que aceptamos la hipótesis nula que indica que el modelo es el más adecuado considerando las tres covariables ($Edad$, $Alta$ y $Flag - Saldo$). Con lo que respecta al test de Wald, el $p - value$ prueba que la hipótesis nula que indica que los coeficientes de las tres covariables sean cero debe ser rechazada. Finalmente, la prueba Score al tener un $p - value$ menor al 0.05 nos prueba que la tasa de riesgo para los trabajadores no es la misma, lo que nos lleva verificar que el supuesto de riesgos proporcionales se cumpla.

Verificación del supuesto de riesgos proporcionales

Para comprobar la hipótesis, utilizamos el software de R y los resultados obtenidos se muestran en la tabla de la Figura 4.6

Covariable	p-value
Edad	0.42
Alta	0.73
Saldo	6.20E-05
Global	6.20E-04

Figura 4.6: Verificación de riesgos proporcionales.

De manera que si consideramos las siguientes hipótesis:

$$H_0: \ Se \ satisface \ el \ supuesto \ de \ riesgos \ proporcionales$$

$$vs$$

$$H_1: \ No \ se \ satisface \ el \ supuesto \ de \ riesgos \ proporcionales$$

para el caso de las covariables *Edad* y *Alta* no existe evidencia significativa del 5 % para suponer que no se cumple el supuesto de riesgos proporcionales del modelo de Cox, por lo tanto, la hipótesis nula es aceptada, sin embargo, para la variable $Flag - Saldo$ la hipótesis aceptada es la alternativa lo que nos dice que esta variable viola el supuesto y al mismo tiempo nos lleva a pensar que dentro de los tiempos existe un punto donde la tasa de riesgo cambia para los trabajadores.

Lo mencionado anteriormente también se puede justificar con las graficación de los residuos de Schoenfeld, en la Figura 4.7 se observa que para algunos puntos el del Flag-Saldo el error se aleja de cero, es decir, se corrobora el rechazo de la hipótesis nula y se da por hecho que se viola el supuesto de riesgos proporcionales.

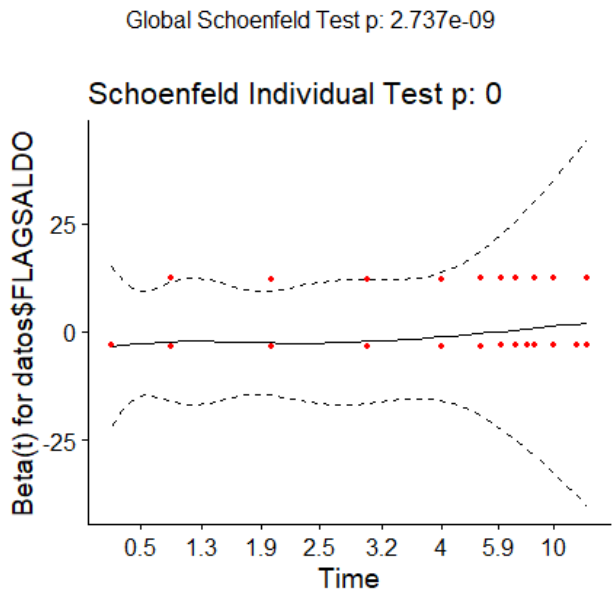

Figura 4.7: Residuos de Schoenfeld para la Covariable Saldo.

4.3. Aplicación del modelo de Cox con punto de cambio

Como en la sección anterior se comprobó que al generar el modelo de Cox se violó el supuesto de riesgos proporcionales el objetivo de esta sección es estimar el punto de cambio que divide a la función de riesgo en dos partes, de manera que si encontramos el punto de cambio y analizamos el supuesto de riesgos proporcionales en cada parte de la función, dicho supuesto se satisfacerá.

Para hallar el punto de cambio, usaremos al igual que en la sección anterior R y el proceso consta en encontrar el punto de cambio a través del método de ensayo y error, es decir, supondremos que el primer tiempo es el punto de cambio, en caso afirmativo podrá corroborarse el supuesto y detendremos el proceso en caso negativo tomaremos el punto de cambio como el segundo tiempo de falla y así sucesivamente hasta encontrarlo.

De manera que si consideramos únicamente la variable que violó el supuesto, lo que buscamos estimar es la siguiente función:

$$h(t|Z) = \begin{cases} h_0(t)\ \exp\left\{\beta_1' Z_1\right\}, & \text{para } t \leq \tau, \\ h_0(t)\ \exp\left\{(\beta_1 + \theta_1)' Z_1\right\}, & \text{para } t > \tau. \end{cases}$$

$$h(t|Z) = \begin{cases} h_0(t)\ \exp\left\{\beta' Flag - Saldo\right\}, & \text{para } t \leq \tau, \\ h_0(t)\ \exp\left\{(\beta + \theta)' Flag - Saldo\right\}, & \text{para } t > \tau. \end{cases}$$

Así entonces con ayuda de R, obtenemos los siguientes resultados:

Valor de T	Estimación de β	Estimación de θ	p-value antes de T	p-value después de T
1	-1.8699	-1.1901	0.0220	0.0000
2	-2.6378	-0.9251	0.0120	0.0000
3	-2.6838	-1.6338	0.3600	0.0290
4	-2.6803	2.2273	0.2100	0.7000

Figura 4.8: Puntos de cambio.

De la Figura 4.8 se puede observar que al ir moviendo el valor de τ encontramos la partición donde el supuesto de riesgos proporcionales se satisface, por ejemplo, si consideramos $\tau = 1$ es claro que podemos estimar los valores de β y θ, sin embargo, al comprobar la proporcionalidad en los riesgos verificamos que el punto de cambio declarado en $\tau = 1$ el supuesto no se cumple, es decir, rechazamos la hipótesis nula que comprueba los riesgos proporcionales, un caso similar pasa cuando tomamos $\tau = 2$ ó $\tau = 3$, sin embargo, cuando declaramos a $\tau = 4$ no existe evidencia suficiente para rechazar la hipótesis nula de riesgos proporcionales, lo que nos lleva a terminar el proceso y poder considerar 4 como el punto de cambio que al particionar los tiempos de falla en dos conjuntos satisface el supuesto principal del modelo de Cox.

De manera que la función de riesgo de modelo de Cox con punto de cambio queda determinada por:

$$h(t|Z) = \begin{cases} h_0(t)\ \exp\left\{\beta' Flag - Saldo\right\}, & \text{para } t \leq \tau, \\ h_0(t)\ \exp\left\{(\beta+\theta)' Flag - Saldo\right\}, & \text{para } t > \tau. \end{cases}$$

$$h(t|Z) = \begin{cases} h_0(t)\ \exp\left\{-2.6803' Flag - Saldo\right\}, & \text{para } t \leq 4, \\ h_0(t)\ \exp\left\{(-2.6803+2.2273)' Flag - Saldo\right\}, & \text{para } t > 4. \end{cases}$$

$$h(t|Z) = \begin{cases} h_0(t)\ \exp\left\{-2.6803 Flag - Saldo\right\}, & \text{para } t \leq 4, \\ h_0(t) \exp\left\{-0.4530 Flag - Saldo\right\}, & \text{para } t > 4. \end{cases}$$

Finalmente, podemos declarar que los trabajadores que tienen un tiempo menor o igual a 4 años tienen un riesgo menor de solicitar su traspaso comparado con aquellos trabajadores que ya tienen más de 4 años dentro de la empresa, lo cual es bastante cierto ya que al tener una antigüedad mucho mayor teniendo un saldo creciente y distinto de cero vuelve a los trabajadores más atractivos para una AFORE competencia.

Conclusiones

En esta investigación el objetivo fue aplicar el modelo de Cox con un punto de cambio a una base de datos reales con el fin de analizar la función de riesgo que tienen elementos de una población, esta base datos concentra los registros de algunos trabajadores afiliados a una empresa de AFORE, por lo que el objeto de estudio fue analizar el riesgo que tienen los trabajadores para cambiar de administradora o en el caso de la empresa traspasar la cuenta de ahorro individual del trabajador a una AFORE competencia incluyendo el efecto de algunas características asociadas a ellos.

En primer lugar se aplicó el modelo de Cox tradicional cuya naturaleza relaciona la función de riesgo de los elementos de estudio a una variable de salida y a un conjunto de variables explicativas, donde estás últimas representan un efecto adicional al riesgo al que se exponen los elementos de estudio. Este modelo considera que existen riesgos proporcionales, es decir si se calcula la razón de riesgo entre dos elementos este será una constante independiente del tiempo. Para calcular el efecto de las covariables se estimó el valor de las β_i que las acompañan a través de las observaciones que se tienen disponibles. Una vez generado el modelo de Cox se analizó si el modelo junto con su suposición ajustaban bien a los datos, sin embargo, no lo fue, y a raíz de eso se procedió a aplicar/generar el modelo de Cox con un punto de cambio.

El modelo de Cox un punto de cambio es una extensión del modelo tradicional, cuya función principal es modelar a un conjunto de datos donde alguna o algunas covariables son dependientes del tiempo, es decir, si se obtiene la razón de riesgos proporcionales entre dos elementos provenientes de niveles diferentes, esta razón será función del tiempo. En este caso la variable donde se violó el supuesto de proporcionalidad fue el $Flag - Saldo$, lo que nos dice que entre mayor sea el tiempo el efecto del saldo en el riesgo de que el trabajador sea traspasado a otra AFORE será mayor, lo que es convincente ya que al pasar el tiempo, si el trabajador genera aportaciones directamente su saldo será distinto de cero lo cual lo hace atractivo para ser llevado o atraído por una AFORE competencia.

Para finalizar es importante mencionar que así como en esta investigación el conocimiento de supervivencia fue de gran utilidad también lo puede ser en cualquier otra rama de estudio,

lo que nos permite tomar decisiones más acertadas para mejorar el negocio o situación.

Para un trabajo futuro es encontrar otra base de datos que permita aplicar el modelo de Cox.

Bibliografía

[1] Cox, D.C., *Analysis od Survival Data*, Chapman and Hall, 1998.

[2] Cunningham, Robin J., *Models for Quantifying Risk*, ACTEX Publications, 2012.

[3] Lee, Elisa T., *Statistical Methods for Survival Data Analysis*, John Wiley and Sons, Inc., Hoboken, 2003.

[4] Melody S., *Survival Analysis with Change Point Hazard Functions*, Revista Harvard University Biostatistics, 2006.

[5] Mendenhall W., Wackerly Dennis D., *Introduction to Probability and Statistics*, Cengage Learning Editores, 2010.

[6] Meyer, Paul L., *Probabilidad y Estadística*, Addison - Wesley Iberoamericana, 1970.

[7] Zimmermann, G., *From Basic Survival Analytic Theory to a Non - Standard Application*, Springer Spektrum, 2017.

Apéndice A

Código

```
library(survival)
library(KMsurv)
library(survMisc)
library(survminer)
library(flexsurv)
library(actuar)
library(dplyr)
library(tidyr)
library(readr)
library(ggplot2)
datos <- read_csv("FORMATO/Nueva carpeta/FINAL.csv")
datos1<- filter(datos, SALDO>0)
datos1$FLAGAPORP<- 1
datos2<- filter(datos, SALDO==0)
datos2$FLAGAPORP<- 0
datos <-rbind(datos1,datos2)
datos1<- filter(datos, APORP>0)
datos1$FLAGSALDO<- 1
datos2<- filter(datos, APORP==0)
datos2$FLAGSALDO<- 0
datos <-rbind(datos1,datos2)
datos<- select(datos,'TIEMPO','CENSURA','GENERO','EDADSEFUE','ALTA','FLAGSALDO'
                    ,'FLAGAPORP','RETIROS')
datos<-datos[order(datos$TIEMPO),]
##MODELO CONSIDERANDO TODAS LAS VARIBALES DISPONIBLES
modelotodasvaribales<- coxph(Surv(TIEMPO,CENSURA) ~ datos$GENERO
                             + datos$EDADSEFUE+datos$ALTA
                             +datos$FLAGSALDO+datos$FLAGAPORP, data=datos)
summary(modelotodasvaribales)
rprop = cox.zph(modelotodasvaribales)
print(rprop)

##MODELO CON LAS VARIBALES MÁS SIGNIFICATIVAS
modelosignificativas<- coxph(Surv(TIEMPO,CENSURA) ~ datos$EDADSEFUE+datos$ALTA
                             +datos$FLAGSALDO, data=datos)
summary(modelosignificativas)
rprop = cox.zph(modelosignificativas)
print(rprop)
ggcoxzph(rprop)
```

```
##MODELO SOLO CON EL FLAG SALDO MODELO DE COX
modeloSALDO<- coxph(Surv(TIEMPO,CENSURA) ~ datos$FLAGSALDO, data=datos)
summary(modeloSALDO)
rprop = cox.zph(modeloSALDO)
print(rprop)
ggcoxzph(rprop)

datos1<- filter(datos, TIEMPO<=1)
ED11<- coxph(Surv(TIEMPO,CENSURA) ~ datos1$FLAGSALDO, data=datos1)
summary(ED11)
rprop1 = cox.zph(ED11)
print(rprop1)
ggcoxzph(rprop1)
datos2<- filter(datos, TIEMPO>1)
ED12<- coxph(Surv(TIEMPO,CENSURA) ~ datos2$FLAGSALDO, data=datos2)
summary(ED12)
rprop2 = cox.zph(ED12)
print(rprop2)
ggcoxzph(rprop2)

datos1<- filter(datos, TIEMPO<=2)
ED11<- coxph(Surv(TIEMPO,CENSURA) ~ datos1$FLAGSALDO, data=datos1)
summary(ED11)
rprop1 = cox.zph(ED11)
print(rprop1)
ggcoxzph(rprop1)
datos2<- filter(datos, TIEMPO>2)
ED12<- coxph(Surv(TIEMPO,CENSURA) ~ datos2$FLAGSALDO, data=datos2)
summary(ED12)
rprop2 = cox.zph(ED12)
print(rprop2)
ggcoxzph(rprop2)

datos1<- filter(datos, TIEMPO<=3)
ED11<- coxph(Surv(TIEMPO,CENSURA) ~ datos1$FLAGSALDO, data=datos1)
summary(ED11)
rprop1 = cox.zph(ED11)
print(rprop1)
ggcoxzph(rprop1)
datos2<- filter(datos, TIEMPO>3)
ED12<- coxph(Surv(TIEMPO,CENSURA) ~ datos2$FLAGSALDO, data=datos2)
summary(ED12)
rprop2 = cox.zph(ED12)
print(rprop2)
ggcoxzph(rprop2)

datos1<- filter(datos, TIEMPO<=4)
ED11<- coxph(Surv(TIEMPO,CENSURA) ~ datos1$FLAGSALDO, data=datos1)
summary(ED11)
rprop1 = cox.zph(ED11)
print(rprop1)
ggcoxzph(rprop1)

datos2<- filter(datos, TIEMPO>4)
ED12<- coxph(Surv(TIEMPO,CENSURA) ~ datos2$FLAGSALDO, data=datos2)
summary(ED12)
rprop2 = cox.zph(ED12)
print(rprop2)
ggcoxzph(rprop2)
```