Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias Físico Matemáticas

Postgrado en Ciencias Matemáticas

Análisis de la deserción en las licenciaturas de la FCFM-BUAP mediante modelos de supervivencia

Tesis presentada para obtener el título de:
Maestra en Ciencias Matemáticas

Presenta:
Lic. Blanca Xochilt Muñoz Vargas

Directores:

Dr. Bulmaro Juárez Hernández
Dra. Lucía Cervantes Gómez

Puebla, Puebla, Octubre 2017

DRA. LIDIA AURORA HERNÁNDEZ REBOLLAR
SECRETARIA DE INVESTIGACIÓN Y
ESTUDIOS DE POSTGRADO, FCFM-BUAP
P R E S E N T E:

Por este medio le informo que el(la) C:

BLANCA XOCHILT MUÑOZ VARGAS

estudiante de la Maestría en Ciencias (Matemáticas), ha cumplido con las indicaciones que el Jurado le señaló en el Coloquio que se realizó el día 18 de septiembre de 2017, con la tesis titulada:

Análisis de la deserción en las licenciaturas de la FCFM-BUAP mediante modelos de supervivencia

Por lo que se le autoriza a proceder con los trámites y realizar el examen de grado en la fecha que se le asigne.

A T E N T A M E N T E.
H. Puebla de Z. a 19 de septiembre de 2017.

DR. FERNANDO MACÍAS ROMERO
COORDINADOR DEL POSTGRADO
EN MATEMÁTICAS.

Ccp.Archivo.
DRA. LAHR/mtrv

Facultad
de Ciencias
Físico Matemáticas

Av. San Claudio y 18 sur, edif. 111 A,
Ciudad Universitaria, Col. San
Manuel, Puebla, Pue. C.P. 72570
01 (222) 229 55 00 Ext. 7550 y 7552

Dedico esta Tesis a:

Mi mamá

Sra. Catalina Vargas Paredes

Mis hermanos

Angel Felipe Muñoz Vargas
Juan Carlos Muñoz Vargas

Mis sobrinas

Kathia Yamilet Muñoz Galicia
Danna Jatzibe Muñoz Galicia

Rafael Pérez Flores

Karen G. Tamayo Pérez
Ana Luisa Nieto Méndez

Agradecimientos

Expreso mi profundo agradecimiento a las siguientes Instituciones y especialistas por el apoyo otorgado:

Facultad de Ciencias Físico Matemáticas de la BUAP por las facilidades y el apoyo para realizar mis estudios de Maestría en Ciencias Matemáticas.

Consejo Nacional de Ciencia y Tecnología (Conacyt) por el apoyo económico brindado durante la maestría.

Vicerrectoría de Investigación y Estudios de Posgrado (VIEP) de la BUAP por las becas otorgadas como apoyo a la movilidad estudiantil.

Dra. Lucía Cervantes Gómez y Dr. Bulmaro Juárez Hernández, por todas sus enseñanzas, su apoyo y el tiempo dedicado a la dirección de esta tesis.

Dr. Víctor Hugo Vázquez Guevara, Dra. Hortensia Josefina Reyes Cervantes, Dra. Gladys Linares Fleites y M.C. Julio Erasto Poisot Macías por su dedicación en la revisión de esta tesis, sus correcciones y contribuciones.

Dra. Olga Leticia Fuchs Gómez por la información facilitada, la cual se utilizó en el análisis del caso de estudió de esta tesis, su apoyo y disposición para aclarar dudas.

M.C. Edgar Santiago Moyotl Hernández por su ayuda en la organización de la información del caso de estudió de esta tesis.

Introducción

La deserción estudiantil universitaria ha sido estudiada por diferentes autores e Instituciones de Educación Superior. Este tema ha tomado un lugar importante dado que no tiene sentido realizar un esfuerzo significativo por aumentar la cobertura, calidad y equidad en educación superior, sin controlar la deserción y su problemática multicausal y compleja [28]. Las licenciaturas de la Facultad de Ciencias Físico Matemáticas (FCFM) de la Benemérita Universidad Autónoma de Puebla (BUAP) no están excentas del problema de deserción, hay alumnos desertores en los diferentes semestres impartidos. Además, desde el 2006 hasta el 2013 el porcentaje de deserción en el primer año de la Licenciatura en Matemáticas y la Licenciatura en Matemáticas Aplicadas ha variado entre el 26 % y el 57 %, mientras que en el caso de la Licenciatura en Actuaría desde el 2010 hasta el 2013 la deserción varía entre el 13 % y el 25 %.

Los estudios acerca de la deserción son importantes, debido a que permiten comprender mejor el problema, lo cual es importante para generar estrategias de retención estudiantil. En la FCFM se han realizado diferentes estudios en los cuales se investigan algunos de los problemas relacionados con aprovechamiento, aprobación, reprobación, egreso y titulación, con el fin de comprender mejor la situación y poder generar estrategias adecuadas ([3], [12], [15], [26], [29]). En este trabajo se realiza un análisis de la deserción en las licenciaturas de la FCFM-BUAP mediante modelos de supervivencia. El Análisis de Supervivencia centra su interés en el tiempo hasta que ocurre cierto evento específico, frecuentemente llamado falla, y sus aplicaciones van desde investigaciones de la durabilidad de artículos manufacturados hasta estudios de enfermedades humanas y sus tratamientos. Cabe mencionar que esta metodología ha sido utilizada en estudios de deserción escolar en universidades de otros países [28], sin embargo, no se tiene el conocimiento de que ya se haya utilizado en México.

Antes de realizar el análisis de la deserción en la FCFM-BUAP se revisaron los conceptos básicos del Análisis de Supervivencia y diferentes modelos de supervivencia: paramétricos, no paramétricos y semiparamétricos, algunos modelos que sólo involucran a la variable tiempo de falla y otros que también involucran variables explicativas.

La primera parte del análisis del tiempo de deserción de los alumnos de las licenciaturas de la FCFM se realiza mediante modelos no paramétricos y

modelos paramétricos. Estos modelos permiten hacer comparaciones de la estimación de la función de supervivencia, de la estimación de la función de riesgo y también de la estimación de la función de riesgo acumulado de las diferentes licenciaturas impartidas en la FCFM.

La segunda parte del análisis del tiempo de deserción se realiza mediante el modelo de riesgo proporcional semiparamétrico, debido a que este modelo está enfocado en evaluar la relación con las covariables, por lo que proporciona factores indicadores de mayor riesgo de deserción. Los factores considerados en este estudio son: puntaje de ingreso, autoestima, hábitos de estudio, razonamiento científico, comprensión lectora , estilos de aprendizaje, género, con quien vive, financiamiento del bachillerato de procedencia, tipo de bachillerato de procedencia, materias reprobadas en el bachillerato, opción de carrera, sostén de estudios, trabajo y recursos semanales.

El objetivo principal del trabajo es analizar la deserción de los estudiantes de las carreras de Actuaría, Matemáticas y Matemáticas Aplicadas de la FCFM-BUAP para identificar los períodos y las características de los alumnos que implican un mayor riesgo de deserción. El estudio de deserción se hace mediante diferentes tipos de modelos del Análisis de Supervivencia ya que esta metodología es muy útil para este tipo de estudios y permite procesar la base de datos de deserción de la FCFM.

El presente trabajo está organizado en 4 capítulos, la conclusión y 5 apéndices. En el Capítulo 1 se presentan los conceptos básicos del Análisis de Supervivencia, en él se presentan los requisitos para determinar con precisión el tiempo de falla (Sección 1.1), la definición de función de supervivencia y de la función de riesgo, además de sus características y las relaciones que existen entre ellas (Sección 1.2) y algunas categorías de censura y truncamiento (Sección 1.3).

El Capítulo 2 contiene modelos de supervivencia: paramétricos y no paramétricos. En él se calcula la función de verosimilitud de datos censurados por la derecha (Sección 2.1) y de datos censurados por intervalo (Sección 2.2); además, se presenta inferencia basada en la función de verosimilitud (Sección 2.3). Se describen algunas familias paramétricas que se usan como modelos en el análisis de datos de tiempo de fallas (Sección 2.4). También se proporcionan modelos no paramétricos (Sección 2.5), el Estimador Kaplan-Meier y el Estimador de Nelson-Aalen. Por último, se presenta el Criterio de información de Akaike y el Criterio de información de Bayesiano (Sección 2.6) que son herramientas para la selección de modelos.

El Capítulo 3 muestra modelos de regresión de supervivencia (paramétricos y semiparamétricos) los cuales se usan para representar el efecto de

variables explicativas o covariables en el tiempo de falla. Los modelos paramétricos presentados son: Modelo de regresión exponencial (Subsección 3.1.2), Modelo de regresión Weibull (Subsección 3.1.3), Modelo de vida acelerada (Subsección 3.1.4), Modelo de riesgo proporcional (Subsección 3.1.5) y Modelo de riesgo aditivo (Subsección 3.1.6). El modelo semiparamétrico presentado es el Modelo de riesgo proporcional semiparamétrico (Sección 3.2). También se presenta la función de verosimilitud para datos que contienen covariables y están sujetos a censura por la derecha o por intervalo (Subsección 3.1.1).

En el Capítulo 4 se analiza la deserción en las licenciaturas de la FCFM de la BUAP mediante algunos de los modelos presentados en los Capítulos 2 y 3.

En los Apéndices A y B se muestran algunos resultados estadísticos y matemáticos, respectivamente, con los cuales se desarrolla el Capítulo 2. En el Apéndice C se describe la regresión paso a paso (*stepwise*) que es una técnica de selección de modelos ampliamente usada en la regresión lineal múltiple. El Apéndice D proporciona una lista de comandos de R con los cuales se lleva acabo el análisis del Capítulo 4. Por último, en el Apéndice E se presentan algunas de las salidas del *software* R obtenidas en el estudio del Capítulo 4.

Índice general

Capítulo 1

Conceptos básicos

En el Análisis de Supervivencia, el interés se centra en un grupo o grupos de individuos para cada uno de los cuales se define un evento específico, frecuentemente llamado fracaso, falla o muerte, que ocurre después de un periodo de tiempo llamado el tiempo de falla, tiempo de supervivencia o tiempo de vida. El fracaso puede ocurrir como máximo una vez en cada individuo [9]. Las aplicaciones de la metodología de la distribución del tiempo de falla van desde investigaciones de la durabilidad de artículos manufacturados hasta estudios de enfermedades humanas y sus tratamientos [25].

Algunas veces se tiene interés solamente en la distribución del tiempo de falla de un solo grupo. Más a menudo, se desea comparar los tiempos de falla de dos o más grupos para ver, por ejemplo, si el tiempo de falla de los individuos son sistemáticamente más largos en el segundo grupo que en el primero. Alternativamente, pueden estar disponibles para cada individuo valores de variables explicativas, consideradas relacionadas a la supervivencia [9].

El Análisis de Supervivencia se considera propiamente una técnica univariada más que una técnica multivariada ya que sólo tiene una variable respuesta, el tiempo de falla, aunque haya muchas variables explicativas.

1.1. El tiempo de falla

Para determinar con precisión el tiempo de falla, hay tres requisitos: el tiempo origen debe estar bien definido, se debe acordar una escala para medir el paso del tiempo y finalmente el significado de fracaso debe ser del todo claro.

Además de que el tiempo origen debe definirse con precisión para cada

individuo, también es deseable que, sujeto a cualquier diferencia conocida en las variables explicativas, todos los individuos deben ser comparables como sea posible en el tiempo origen. El tiempo origen no tiene que ser y por lo general no es la misma fecha del calendario para cada individuo, por lo que, el tiempo de falla de cada individuo se mide desde su propia fecha de entrada. El tiempo origen no necesariamente es el punto en que un individuo entra al estudio, si es así, se necesitan métodos especiales y se hace referencia a tales datos como truncados por la izquierda.

Frecuentemente la escala para medir el tiempo es el tiempo del reloj (tiempo real), aunque surgen otras posibilidades, tales como el uso de tiempo de funcionamiento de un sistema, kilometraje de un carro, o alguna medida de la carga acumulada encontrada. De hecho, en muchas aplicaciones industriales de fiabilidad, es más apropiado medir el tiempo por el uso acumulativo, en algún sentido. El único requisito universal para los tiempos de falla es que sean no negativos.

Una de las razones para la elección de una escala de tiempo es el significado directo para el individuo en cuestión, que justifica el uso del tiempo real en la investigación de la supervivencia en un contexto médico. Otra consideración es que dos individuos tratados del mismo modo deben, en igualdad de circunstancias, estar en un estado similar al cabo de tiempos iguales; esta es la base para el uso de la carga acumulada alentada en un contexto de ingeniería. Si dos o más formas diferentes de medir el tiempo están disponibles, puede ser posible, después de haber seleccionado la escala de tiempo más apropiada, usar otros tiempos como variables explicativas.

Por último, el significado del evento específico de falla se debe definir con precisión. En el trabajo médico, el fracaso podría significar la muerte, la muerte por una causa específica, la primera reaparición de una enfermedad después del tratamiento, o la incidencia de una enfermedad nueva. En algunas aplicaciones hay poco o nada de arbitrariedad en la definición de fracaso. En otras, por ejemplo en algunos contextos industriales, falla se define como la primera instancia en que el rendimiento, medido de alguna manera cuantitativa, cae por debajo de un nivel aceptable, definido tal vez por una especificación. Entonces habrá algunas arbitrariedades en la definición de falla y esto se considerará para concentrarse en que se analizará el tiempo de falla o se analizará la medida de rendimiento total como una función del tiempo [9].

1.2. Definiciones

Se considera una población homogénea de individuos, teniendo cada uno un tiempo de falla. Es decir, se trata con una variable aleatoria no negativa, T. En particular, se supone que están definidos claramente un tiempo origen y una escala de medición de tiempo.

1.2.1. Distribuciones continuas

Primero, supóngase que T es continua. Sean $f(\cdot)$ la función de densidad de probabilidad (f.d.p.) y

$$F(t) = Pr(T \leqslant t) = \int_0^t f(x)dx, \text{ para } t \geqslant 0,$$

la función de distribución acumulada (f.d.a.) de T.

La probabilidad de que un individuo sobreviva al tiempo t está dada por la función de supervivencia

$$S(t) = Pr(T > t) = \int_t^\infty f(x)dx = 1 - F(t), \tag{1.1}$$

por lo que

$$f(t) = -S'(t). \tag{1.2}$$

En algunos contextos que implican sistemas o vidas de artículos fabricados, $S(\cdot)$ se conoce como la función de fiabilidad.

La función $S(\cdot)$ es continua, no creciente con $S(0) = 1$ y $\lim_{t\to\infty} S(t) = 0$. En ocasiones, se puede permitir que $\lim_{t\to\infty} S(t) > 0$ para considerar ajustes en donde algunos individuos nunca fallan, estos se deben tratar como casos especiales.

El p-ésimo cuantil de la distribución de T es el valor t_p que cumple que

$$F(t_p) = Pr(T \leqslant t_p) = p.$$

El p-ésimo cuantil también se conoce como el $100\ p$-ésimo percentil de la distribución. El 0.5 cuantil se llama la mediana de la distribución.

Las funciones $S(\cdot)$, $f(\cdot)$ y $F(\cdot)$ proporcionan tres formas matemáticas equivalentes para especificar la distribución de una variable aleatoria continua no negativa, y hay por supuesto muchas otras funciones equivalentes. Una con valor especial en el contexto presente es la función de riesgo, definida por,

$$h(t) = \lim_{\Delta\to 0^+} \frac{Pr(t \leqslant T < t + \Delta | t \leqslant T)}{\Delta}. \tag{1.3}$$

La función de riesgo especifica la tasa instantánea de muerte o falla en el tiempo t, dado que el individuo sobrevive hasta el tiempo t; $h(t)\Delta$ es la probabilidad aproximada de falla en $[t, t+\Delta)$, dado que sobrevivió hasta el tiempo t.

De la definición de probabilidad condicional, se tiene que,

$$\begin{aligned} h(t) &= \lim_{\Delta\to 0^+} \frac{Pr(t \leqslant T < t+\Delta, t \leqslant T)}{\Delta Pr(t \leqslant T)} \\ &= \frac{1}{Pr(t \leqslant T)} \lim_{\Delta\to 0^+} \frac{Pr(t \leqslant T < t+\Delta)}{\Delta} \\ &= \frac{1}{S(t)} \lim_{\Delta\to 0^+} \frac{F(t+\Delta) - F(t)}{\Delta} \\ &= \frac{1}{S(t)} F'(t), \end{aligned}$$

i.e.

$$h(t) = \frac{f(t)}{S(t)}. \tag{1.4}$$

Substituyendo (1.2) en (1.4)

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

luego $\log(S(u))|_0^t = -\int_0^t h(u)du$, además como $S(0) = 1$,

$$S(t) = \exp\left(-\int_0^t h(u)du\right) = \exp[-H(t)], \tag{1.5}$$

en donde

$$H(t) = \int_0^t h(u)du \tag{1.6}$$

es llamada la función de riesgo acumulado. Además,

$$f(t) = h(t)\exp[-H(t)]. \tag{1.7}$$

Ocasionalmente, otras representaciones de la distribución del tiempo de falla son útiles. Un ejemplo es la falla residual esperada o vida media residual,

$$r(t) = E(T - t|T \geqslant t), \text{ para } t \geqslant 0,$$

que determina de manera única una distribución de supervivencia continua con media finita [19]. Si T tiene f.d.p. $f(\cdot)$, f.d.a. $F(\cdot)$ y función de supervivencia $S(\cdot)$, entonces $T-t|T \geqslant t$ tiene f.d.p. $f_{T-t|T\geqslant t}(s) = f(s+t)/S(t)$ para $s \in [0,\infty)$. Así

$$r(t) = \frac{1}{S(t)}\int_0^\infty uf(u+t)du = \frac{1}{S(t)}\int_t^\infty (s-t)f(s)ds, \text{ para } t \geqslant 0. \quad (1.8)$$

Notar que de (1.8) se tiene que $r(0) = E(T)$.

La falla residual esperada y la función de supervivencia cumplen que [19]

$$r(t) = \frac{1}{S(t)}\int_t^\infty S(s)ds \quad (1.9)$$

y

$$S(t) = \frac{r(0)}{r(t)}\exp\left\{-\int_0^t \frac{1}{r(u)}du\right\}.$$

Observar que de (1.9) se tiene que

$$E(T) = \int_0^\infty S(s)ds,$$

lo cual también se tiene por el Lema A.2 del Apéndice A.

En la Tabla 1.1 se muestran las diferentes relaciones que existen entre la f.d.p., f.d.a., función de supervivencia, función de riesgo, función de riesgo acumulado y falla residual esperada en el caso de que el tiempo de falla, T, tenga una distribución continua.

1.2.2. Distribuciones discretas

Si T puede ser tratada como una variable aleatoria discreta. Supóngase que T puede tomar los valores $t_0, t_1, t_2, \ldots$ con $0 = t_0 < t_1 < t_2 < \ldots$, y sean la función de probabilidad (f.p.) de T

$$f(t_j) = Pr(T = t_j), j = 0, 1, 2, \ldots.$$

y la f.d.a. de T

$$F(t) = Pr(T \leqslant t) = \sum_{j:t_j \leqslant t} f(t_j).$$

La función de supervivencia es entonces

$$S(t) = Pr(T > t) = \sum_{j:t_j > t} f(t_j) = 1 - F(t).$$

	$f(\cdot)$	$F(\cdot)$	$S(\cdot)$	$h(\cdot)$	$H(\cdot)$	$r(\cdot)$	$f(\cdot)$ y $S(\cdot)$
$f(t) =$		$F'(t)$	$-S'(t)$	$h(t)\exp[-H(t)]$			
$F(t) =$	$\int_0^t f(x)dx$		$1 - S(t)$				
$S(t) =$		$1 - F(t)$			$\exp[-H(t)]$	$\frac{r(0)}{r(t)} \exp\left\{-\int_0^t \frac{1}{r(s)} ds\right\}$	
$h(t) =$			$-\frac{d\log(S(t))}{dt}$				$\frac{f(t)}{S(t)}$
$H(t) =$				$\int_0^t h(u)du$			
$r(t) =$			$\frac{\int_t^\infty S(s)ds}{S(t)}$				

Tabla 1.1: Relaciones entre la f.d.p. ($f(\cdot)$), f.d.a. ($F(\cdot)$), función de supervivencia ($S(\cdot)$), función de riesgo ($h(\cdot)$), función de riesgo acumulado ($H(\cdot)$) y falla residual esperada ($r(\cdot)$).

Cuando es considerada como una función para toda $t \geqslant 0$, $S(\cdot)$ es continua por la derecha, escalonada, no creciente, con $S(0) = 1$ y $\lim_{t\to\infty} S(t) = 0$.

La función de riesgo en tiempo discreto se define como

$$h(t_j) = Pr(T = t_j | T \geqslant t_j) = \frac{Pr(T = t_j)}{Pr(T \geqslant t_j)} = \frac{f(t_j)}{S(t_{j-1})}, j = 0, 1, 2, \ldots, \tag{1.10}$$

donde $S(t_{-1}) = 1$.

Como en el caso continuo, las funciones de probabilidad, supervivencia y riesgo dan especificaciones equivalentes de la distribución de T. Dado que $f(t_j) = S(t_{j-1}) - S(t_j)$, (1.10) implica que

$$h(t_j) = 1 - \frac{S(t_j)}{S(t_{j-1})}, j = 0, 1, 2, \ldots. \tag{1.11}$$

Dado que $S(t_0) = 1$, si $t \in [t_k, t_{k+1})$

$$\begin{aligned} \prod_{j:t_j \leqslant t} (1 - h(t_j)) &= \prod_{j:t_j \leqslant t} \frac{S(t_j)}{S(t_{j-1})} \\ &= \frac{S(t_k)}{S(t_0)} \\ &= S(t_k), \end{aligned}$$

además como $S(\cdot)$ es escalonada y continua por la derecha se tiene que $S(t) = S(t_k)$ por lo que

$$S(t) = \prod_{j:t_j \leqslant t} (1 - h(t_j)). \tag{1.12}$$

Un análogo de $H(\cdot)$, para el caso continuo, puede definirse de dos formas en el caso discreto. Una es por analogía de (1.5), como

$$-\log(S(t)) = -\sum_{j:t_j \leqslant t} \log(1 - h(t_j)).$$

Sin embargo, ésta no es igual a

$$\sum_{j:t_j \leqslant t} h(t_j),$$

que es el segundo análogo [25].

En la Tabla 1.2 se muestran las diferentes relaciones que existen entre la f.p., f.d.a., función de supervivencia y función de riesgo en caso de que el tiempo de falla, T, tenga una distribución discreta.

	$f(\cdot)$	$F(\cdot)$	$S(\cdot)$	$h(\cdot)$	$f(\cdot)$ y $S(\cdot)$
$f(t_j) =$		$F(t_j) - F(t_{j-1})$	$S(t_{j-1}) - S(t_j)$		
$F(t) =$	$\sum_{j:t_j \leqslant t} f(t_j)$				
$S(t) =$	$\sum_{j:t_j > t} f(t_j)$	$1 - F(t)$		$\prod_{j:t_j \leqslant t} (1 - h(t_j))$	
$h(t_j) =$			$1 - \frac{S(t_j)}{S(t_{j-1})}$		$\frac{f(t_j)}{S(t_{j-1})}$

Tabla 1.2: Relaciones entre la f.p. ($f(\cdot)$), f.d.a. ($F(\cdot)$), función de supervivencia ($S(\cdot)$) y función de riesgo ($h(\cdot)$).

1.2.3. La función de riesgo

La función de riesgo es una característica particularmente importante de una distribución de tiempo de falla. Ésta indica la forma en que el riesgo de fallar varía en el tiempo y esto es de interés en la mayoría de las aplicaciones. La información previa sobre la forma de la función de riesgo puede ayudar a guiar la selección del modelo.

Algunas de las posibles formas de la función de riesgo son:

a) Constante: Las funciones de riesgo aproximadamente constantes tienden a ocurrir en entornos estables donde la falla o muerte se debe a fenómenos aleatorios tales como choques o accidentes, que son externos al individuo.

b) Creciente: Las distribuciones con función de riesgo creciente se observan en los individuos para los cuales ocurre algún tipo de envejecimiento o desgaste.

c) Decreciente: Ciertos tipos de dispositivos electrónicos muestran una función de riesgo decreciente cuando los artículos con defectos fallan y se eliminan de la población.

d) Curva de bañera: La función de riesgo es llamada curva de bañera si primero decrece hasta un mínimo y luego crece. Las poblaciones que muestran una función de riesgo en forma de curva de bañera a veces son limpiadas de individuos débiles dejando una población reducida con una función de riesgo creciente.

e) Curva de bañera inversa: En este caso, la función de riesgo primero crece hasta un máximo y luego decrece. Esta forma de la función de riesgo se encuentra en muchas aplicaciones, por ejemplo, en el caso de la supervivencia después de un tratamiento de cáncer donde algunos individuos se curan [25].

1.3. Censura y truncamiento

En los estudios de tiempo de falla puede haber limitaciones en la información recolectada, éstas pueden ser impuestas por el tiempo, costo y otras restricciones. A continuación se presentan dos características que usualmente presentan los datos de tiempo de falla: la censura y el truncamiento.

1.3.1. Censura: por la derecha, por la izquierda, por intervalo y doble

La censura ocurre cuando se conoce que algunos tiempos de falla han ocurrido en cierto intervalo de tiempo y el resto de los tiempos de falla son conocidos exactamente. Existen varias categorías de censura, a continuación se describe la censura: por la derecha, por la izquierda, por intervalo y doble.

Censura por la derecha

El terminar el seguimiento de un estudio de tiempo de falla antes de que algunos individuos fallen causa que sus tiempos de falla sean censurados por la derecha, es decir, sólo están disponibles los límites inferiores del tiempo de falla. La censura por la derecha puede ocurrir por varias razones, ésta puede ser planeada, como cuando se toma la decisión de terminar una prueba de falla antes de que todos los elementos fallen; o imprevista, como cuando una persona en un estudio prospectivo[1] se "pierde del seguimiento" porque se muda lejos de la región donde se realiza el estudio. A continuación se describen algunos mecanismos de censura por la derecha.

1. Censura tipo I
   Se dice que se aplica un mecanismo de censura tipo I cuando cada individuo tiene un tiempo de censura potencial $C_i > 0$ tal que T_i es observado si $T_i \leqslant C_i$; de otra manera, sólo se sabe que $T_i > C_i$. La censura tipo I frecuentemente surge cuando un estudio se lleva a cabo durante un período de tiempo especificado.

2. Censura aleatoria independiente
   Un proceso de censura aleatoria muy simple que es a menudo realista es aquel en el que cada individuo se asume que tiene un tiempo de falla T y un tiempo de censura C, donde T y C son variables aleatorias continuas independientes, con función de supervivencia $S(\cdot)$ y $G(\cdot)$, respectivamente. Se asume que todos los tiempos de falla y tiempos de censura son mutuamente independientes y además que $G(\cdot)$ no depende de los parámetros de $S(\cdot)$.

3. Censura tipo II
   El término de censura tipo II se refiere a la situación en donde sólo se

[1] En un estudio prospectivo de tiempo de falla, los individuos son seguidos desde el tiempo de entrada hasta el tiempo de falla o censura. En algunos estudios, el plan observacional es retrospectivo en algún grado, es decir, parte o todo el periodo de observación ocurre cronológicamente antes de la selección del individuo [25].

observan los r tiempos de falla más pequeños $t_{(1)} \leqslant t_{(2)} \leqslant \ldots \leqslant t_{(r)}$ en una muestra aleatoria de tamaño n; aquí r es un entero específico entre 1 y n. Este mecanismo de censura surge cuando en el estudio comienzan n individuos en el mismo tiempo, terminando el estudio una vez que se observan r fallas (o tiempos de falla).

4. Censura tipo II progresiva
   La censura tipo II progresiva es una generalización de la censura tipo II. En este caso, se observan las primeras r_1 ($r_1 \leqslant n$) fallas en una prueba de falla de n individuos; luego n_1 ($n_1 \leqslant n - r_1$) de los restantes $n - r_1$ individuos sin falla se retiran del experimento, dejando $n - r_1 - n_1$ individuos aun presentes. Cuando nuevamente fallan r_2 ($r_2 \leqslant n - r_1 - n_1$) individuos, se retiran n_2 ($n_2 \leqslant n - r_1 - n_1 - r_2$) individuos sin falla y así sucesivamente. El experimento termina después de algunas series preestablecidas de repeticiones de este procedimiento.

Suponga que n individuos tienen tiempos de falla representados por las variables aleatorias $T_1, \ldots, T_n$. En lugar de los valores observados para tiempo de falla, se tiene un tiempo t_i que se sabe que es tiempo de falla o tiempo censurado por la derecha. Se define una variable

$$\delta_i = \begin{cases} 1 & \text{si} \quad T_i = t_i \\ 0 & \text{si} \quad T_i > t_i \end{cases},$$

ésta se llama indicador de censura o estado para t_i, dado que indica si t_i es un tiempo de falla observado ($\delta_i = 1$) o tiempo censurado ($\delta_i = 0$). Los datos observados entonces consisten de (t_i, δ_i), $i = 1, \ldots, n$. Con esta notación t_i representa ya sea una variable aleatoria o un valor observado [25].

Censura por la izquierda

Los individuos también pueden estar sujetos a censura por la izquierda, que ocurre si se observa que el individuo falla antes de algún tiempo t, pero el tiempo real de falla es desconocido. En este caso, se observa que $T \in [0, t]$, que es análogo a la censura por la derecha, donde se observa que $T \in (t, \infty]$ [19]. Los datos provenientes de una muestra censurada por la izquierda se pueden representar con (t_i, δ_i), $i = 1, \ldots, n$, donde

$$\delta_i = \begin{cases} 1 & \text{si} \quad T_i = t_i \quad \text{(tiempo de falla observado)} \\ 0 & \text{si} \quad T_i < t_i \quad \text{(tiempo censurado).} \end{cases}$$

Censura por intervalo

En algunos ajustes sólo puede ser posible determinar si el individuo i falla o no en una sucesión de tiempos t_{i1}, t_{i2}, ..., t_{im_i} ($t_{i1} < t_{i2} < \ldots < t_{im_i}$); en este caso, sólo se sabe que el tiempo de falla pertenece a algún intervalo $(t_{i,j-1}, t_{ij}]$, una característica conocida como censura por intervalo o censura intermitente. A continuación se enuncian algunos casos de censura por intervalo.

1. Datos agrupados
   El caso en donde los tiempos de observación son los mismos para todos los individuos (i.e. $t_{ij} = t_j$) son referidos a menudo como datos agrupados.

2. Datos de estado actual
   El término de datos de estado actual se refiere a tiempos de falla censurados por intervalo donde el intervalo para un individuo es ya sea $(0, t_i]$ o (t_i, ∞). Tales datos surgen cuando el individuo i se examina sólo una vez, en el tiempo t_i, en tal punto se determina si la falla ya ocurrió (i.e. $T_i \leqslant t_i$) o no (i.e. $T_i > t_i$).

Los datos provenientes de una muestra censurada por intervalo consisten de un intervalo $(U_i, V_i]$ para cada individuo, donde $U_i < V_i$ y $U_i, V_i \geqslant 0$.

Censura doble

En muchas aplicaciones, el tiempo de falla es el tiempo entre dos eventos, por ejemplo, el tiempo entre la infección con el Virus de la inmunodeficiencia humana (VIH) y el diagnóstico del síndrome de inmunodeficiencia adquirida (SIDA). Si el tiempo del evento inicial está censurado por intervalo, entonces aún si el tiempo exacto de la falla o censura se observa, sólo se sabe que el tiempo de falla o tiempo censurado exacto para T pertenece a un intervalo.

Específicamente, sea U_i^* el tiempo del evento inicial y suponga que sólo se observa que $L_i^* < U_i^* \leqslant R_i^*$ conforme a un esquema que satisface las condiciones especificadas. Sea y_i el tiempo de censura o de falla observado (i.e. el tiempo del segundo evento), medido en la misma escala que U_i^*. Entonces el tiempo de falla o censura para T_i es $t_i = y_i - U_i^*$ y sólo se sabe que $y_i - R_i^* \leqslant t_i < y_i - L_i^*$. Esto es conocido como censura doble o doble censura [25].

1.3.2. Truncamiento

Una segunda característica de muchos estudios de supervivencia, a veces confundida con la censura, es el truncamiento. El truncamiento de los datos de supervivencia se produce cuando sólo se observan aquellos individuos cuyo tiempo de falla se encuentra dentro de un determinado intervalo de observación (Y_L, Y_R), donde $Y_L < Y_R$ y $Y_L, Y_R \geqslant 0$. No se observa a un individuo cuyo tiempo de falla no está en este intervalo y no hay información disponible sobre este individuo para el investigador. Esto contrasta con la censura donde hay al menos información parcial sobre cada individuo. Debido a que sólo se tiene información de los individuos con tiempos de falla en el intervalo de observación, la inferencia de los datos truncados se limita a la estimación condicional.

Cuando $Y_R = \infty$ entonces se tiene truncamiento por la izquierda. Aquí, sólo se observan aquellos individuos cuyo tiempo de falla T excede el tiempo de truncamiento Y_L. Esto es, observamos T si y sólo si $Y_L < T$. Este tiempo de truncamiento se denomina a menudo tiempo de entrada retrasada dado que sólo se observan a los individuos desde este momento hasta que fallan o son censurados. Obsérvese que, a diferencia de la censura por la izquierda en la que se tiene información parcial sobre individuos que experimentan la falla antes del tiempo de ingreso, para el truncamiento por la izquierda estos individuos nunca se considerarán para su inclusión en el estudio.

El truncamiento por la derecha ocurre cuando $Y_L = 0$. Es decir, se observa el tiempo de falla T sólo cuando $T \leqslant Y_R$. Un ejemplo de estudios que presentan truncamiento por la derecha es el estudio de mortalidad basado en registros de muerte [22].

Capítulo 2

Modelos de supervivencia

En el análisis de datos de tiempo de falla se utilizan diferentes tipos de modelos: paramétricos, no paramétricos y semiparamétricos. En los modelos paramétricos se especifica la forma funcional de la distribución que los tiempos de falla tendrían en la ausencia de censura y la inferencia estadística se basa en la metodología de máxima verosimilitud. Para obtener la función de verosimilitud o las propiedades de los procedimientos estadísticos basados en datos censurados es necesario considerar el proceso por el que surgen los tiempos de falla y tiempos censurados. Para hacer esto, aparentemente se necesita un modelo de probabilidad para el mecanismo de censura. De manera interesante, resulta que la función de verosimilitud observada para los parámetros del tiempo de falla toma la misma forma bajo una gran variedad de mecanismos.

2.1. La función de verosimilitud de datos censurados por la derecha

Primero se hallarán las funciones de verosimilitud para una muestra con censura tipo I y tipo II, respectivamente, después para un proceso de censura por la derecha más general con lo cual se llega a que la función de verosimilitud toma la misma forma para los mecanismos de censura considerados.

2.1.1. Datos con censura tipo I

La función de verosimilitud para una muestra de tamaño n con censura tipo I está basada en la distribución de probabilidad de (T_i^*, δ_i), $i = 1, \ldots, n$,

en donde

$$T_i^* = \text{mín}\{T_i, C_i\} \qquad \text{y} \qquad \delta_i = I(T_i \leqslant C_i), \tag{2.1}$$

ambas son variables aleatorias e $I(\cdot)$ es la función indicadora. Dado que las C_i son constantes fijas, T_i^* puede tomar los valores $t \leqslant C_i$ y δ_i sólo toma los valores 0 y 1.

Para $\Delta t \to 0^+$ se cumple que

$$\begin{aligned} Pr(t \leqslant T_i^* < t + \Delta t, \delta_i = 0) &= Pr(t \leqslant T_i^* < t + \Delta t, T_i > C_i) \\ &= Pr(t \leqslant T_i^* < t + \Delta t | T_i > C_i) Pr(T_i > C_i). \end{aligned}$$

Si $t = C_i$ se tiene que

$$\begin{aligned} Pr(t \leqslant T_i^* < t + \Delta t | T_i > C_i) &= Pr(C_i \leqslant \text{mín}\{T_i, C_i\} < C_i + \Delta t | T_i > C_i) \\ &= Pr(C_i \leqslant C_i < C_i + \Delta t) \\ &= 1 \end{aligned}$$

y si $t < C_i$ se sigue que

$$\begin{aligned} Pr(t \leqslant T_i^* < t + \Delta t | T_i > C_i) &= Pr(t \leqslant \text{mín}\{T_i, C_i\} < t + \Delta t | T_i > C_i) \\ &= Pr(t \leqslant C_i < t + \Delta t) \\ &= 0, \end{aligned}$$

para Δt suficientemente pequeño. Por lo tanto

$$\lim_{\Delta t \to 0^+} Pr(t \leqslant T_i^* < t + \Delta t, \delta_i = 0) = \begin{cases} Pr(T_i > C_i) & \text{si} \quad t = C_i \\ 0 & \text{si} \quad t < C_i. \end{cases} \tag{2.2}$$

Por otro lado,

$$\begin{aligned} Pr(t \leqslant T_i^* < t + \Delta t, \delta_i = 1) &= Pr(t \leqslant T_i^* < t + \Delta t | T_i \leqslant C_i) Pr(T_i \leqslant C_i) \\ &= \frac{Pr(t \leqslant T_i < t + \Delta t)}{F(C_i)} F(C_i) \\ &= Pr(t \leqslant T_i < t + \Delta t), \end{aligned}$$

luego

$$\lim_{\Delta t \to 0^+} Pr(t \leqslant T_i^* < t + \Delta t, \delta_i = 1) = \lim_{\Delta t \to 0^+} Pr(t \leqslant T_i < t + \Delta t) = f(t_i). \tag{2.3}$$

De (2.2) y (2.3) se tiene que

$$f_{(T_i^*, \delta_i)}(t, s) = \begin{cases} f(t)^s Pr(T_i > t)^{1-s}, & \text{si} \quad \{t \leqslant C_i \text{ y } s = 1\} \text{ o } (t, s) = (C_i, 0) \\ 0, & \text{si} \quad t < C_i \text{ y } s = 0. \end{cases} \tag{2.4}$$

Asumiendo que los tiempos de falla $T_1, \cdots, T_n$ son estadísticamente independientes, se obtiene la función de verosimilitud de (2.4) como

$$L = \prod_{i=1}^{n} f_{(T_i^*, \delta_i)}(t_i, \delta_i) = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}. \tag{2.5}$$

2.1.2. Datos con censura tipo II

En el caso de censura tipo II se elige a r ($1 \leqslant r \leqslant n$) antes de recolectar los datos y estos consisten de los r tiempos de falla más pequeños en una muestra aleatoria $T_1, \ldots, T_n$. Para distribuciones continuas se pueden denotar los r tiempos de falla más pequeños como $T_{(1)} < T_{(2)} < \ldots < T_{(r)}$. Si T_i tiene f.d.p. $f(\cdot)$, f.d.a. $F(\cdot)$ y función de supervivencia $S(\cdot)$, entonces la f.d.p. conjunta de $T_{(1)}, T_{(2)}, \ldots, T_{(r)}$ es

$$\frac{n!}{(n-r)!} \prod_{i=1}^{r} f(t_{(i)})[1 - F(t_{(r)})]^{n-r} = \frac{n!}{(n-r)!} \left(\prod_{i=1}^{r} f(t_{(i)}) \right) S(t_{(r)})^{n-r}.$$

En términos de la notación (t_i, δ_i) los individuos cuyos tiempos de falla fueron censurados cumplen que $t_i = t_{(r)}$ y $\delta_i = 0$. Luego se tiene que para censura tipo II

$$L \propto \prod_{i=1}^{n} f(t_{(i)})^{\delta_i} S(t_{(r)})^{1-\delta_i},$$

por lo que la verosimilitud para censura tipo I y tipo II es de la misma forma [25].

2.1.3. Datos con censura por la derecha

Ahora se hallará la función de verosimilitud para datos censurados por la derecha de forma más general, primero para el caso de una distribución discreta y luego para el caso de una distribución continua.

Distribuciones discretas

Suponga que n individuos son seguidos desde $t = 0$ hasta que fallan o son censurados. Además, suponga que los tiempos de falla y tiempos censurados son discretos; por conveniencia y sin pérdida de generalidad se asume que los valores para cada uno son $t = 0, 1, 2, \ldots$. Sean $h(\cdot)$ y $S(\cdot)$ la función de riesgo y la función de supervivencia, respectivamente, de T_i, $i = 1, \ldots, n$.

Se introduce la siguiente notación dirigida a la evolución de los procesos de falla y censura sobre el tiempo. Para $t = 0, 1, 2, \ldots$ sean

$$\begin{aligned} Y_i(t) &= I(T_i \geqslant t, \text{El individuo } i \text{ no está censurado antes de } t), \\ dN_i(t) &= Y_i(t) I(T_i = t) \text{ y} \\ dC_i(t) &= Y_i(t) I(\text{El individuo } i \text{ está censurado en } t). \end{aligned}$$

La variable $Y_i(\cdot)$ frecuentemente se llama el indicador de riesgo; éste es igual a 1 si, y sólo si el individuo i no ha fallado y no está censurado antes del tiempo t, y por lo tanto está en riesgo de ser observado fallar en t. Las variables $dN_i(t)$ y $dC_i(t)$ registran las fallas observadas y eventos censurados en el tiempo t, respectivamente. Entre todos los valores $\{dN_i(t), dC_i(t)|t \geqslant 0\}$, sólo uno es distinto de cero para todos los individuos. Estas definiciones se ilustran en la Figura 2.1.

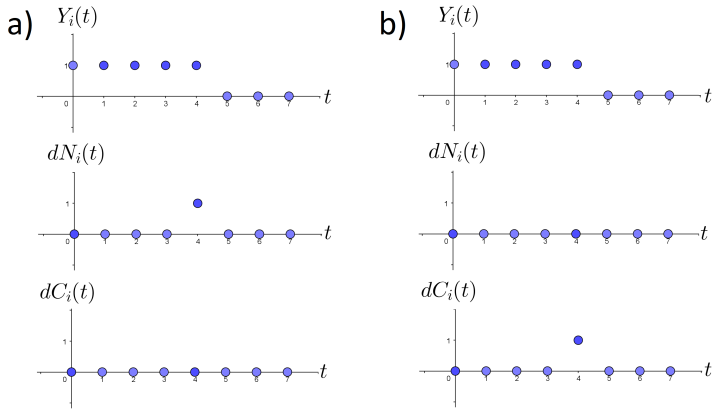

Figura 2.1: Se muestran las gráficas de las funciones $Y_i(\cdot)$, $dN_i(\cdot)$ y $dC_i(\cdot)$, en a) el individuo i tiene tiempo de falla $t_i = 4$ y en b) el individuo i tiene tiempo censurado $t_i = 4$.

También se definen los vectores

$$\begin{aligned} \underline{dN(t)} &= (dN_1(t), \ldots, dN_n(t))', \\ \underline{dC(t)} &= (dC_1(t), \ldots, dC_n(t))' \end{aligned}$$

y

$$\mathcal{H}(t) = \{(\underline{dN(s)}, \underline{dC(s)})|s = 0, 1, \ldots, t-1\}.$$

Se hace referencia a $\mathcal{H}(t)$ como la historia de los procesos de falla y censura en el tiempo t. Ésta consiste de la información sobre todos los eventos de falla y censura que ocurrieron hasta el tiempo $t-1$.

El punto importante es que los $datos$ observados se pueden representar como

$$datos = \{(\underline{dN(t)}, \underline{dC(t)})|t = 0, 1, 2, \ldots\}.$$

Además, se puede descomponer $Pr(datos)$ como

$$Pr(datos) = \prod_{t=0}^{\infty} Pr(\underline{dN(t)}|\mathcal{H}(t)) Pr(\underline{dC(t)}|\underline{dN(t)}, \mathcal{H}(t)), \qquad (2.6)$$

en donde $\mathcal{H}(0)$ es vacío.

Hasta ahora no se han hecho suposiciones sobre el mecanismo de censura, pero para seguir adelante es necesario hacerlas. Los supuestos que han llegado a ser estándar en el análisis de datos de tiempos de falla requieren que

$$Pr(\underline{dN(t)}|\mathcal{H}(t)) = \prod_{i=1}^{n} h(t)^{dN_i(t)}[1-h(t)]^{Y_i(t)(1-dN_i(t))}. \tag{2.7}$$

Efectivamente, esto requiere que dado $\mathcal{H}(t)$, el mecanismo de falla para los individuos en riesgo en el tiempo t opere independientemente y que para $t = 0, 1, 2, \ldots$

$$Pr(dN_i(t) = 1|\mathcal{H}(t)) = Y_i(t)h(t), \tag{2.8}$$

esto significa que la censura en el tiempo t no puede estar relacionada a la información de falla en t o después de t, por lo que no puede discriminar selectivamente entre los individuos de acuerdo a cuando van a fallar en el futuro.

En (2.7) se utiliza la convención de que $0^0 = 1$, correspondiente al hecho de que si $Y_i(t) = 0$ no hay información sobre el individuo i en el tiempo t, y el término en la verosimilitud debe ser igual a 1. Notar que el valor de $Y_i(t)$ está determinado por la información en $\mathcal{H}(t)$.

La condición (2.8) representa una independencia condicionada en $\mathcal{H}(t)$ entre falla y censura en el tiempo t, y los mecanismos que satisfacen esto son llamados frecuentemente mecanismos de censura independiente. Bajo (2.8), la probabilidad de que un individuo que no ha fallado y no está censurado justo antes del tiempo t es observado que falla en t es igual a $h(t)$, lo mismo que sino hubiera censura.

Si los términos $Pr(\underline{dC(t)}|\underline{dN(t)}, \mathcal{H}(t))$ en (2.6) no involucran a ninguno de los parámetros que especifican a $h(t)$, el mecanismo de censura es llamado no informativo; además, estos términos se pueden eliminar de la verosimilitud. Luego se obtiene que

$$\begin{aligned} L &\propto \prod_{t=0}^{\infty} Pr(\underline{dN(t)}|\mathcal{H}(t)) \\ &= \prod_{t=0}^{\infty}\prod_{i=1}^{n} h(t)^{dN_i(t)}[1-h(t)]^{Y_i(t)(1-dN_i(t))}, \end{aligned}$$

por lo que

$$L \propto \prod_{i=1}^{n}\prod_{t=0}^{\infty} h(t)^{dN_i(t)}[1-h(t)]^{Y_i(t)(1-dN_i(t))}, \tag{2.9}$$

o equivalentemente

$$L \propto \prod_{t=0}^{\infty} h(t)^{d_t}[1-h(t)]^{r_t-d_t}, \tag{2.10}$$

en donde $d_t = \sum_{i=1}^{n} dN_i(t)$ y $r_t = \sum_{i=1}^{n} Y_i(t)$ son el número de tiempos de falla observados igual a t y el número de individuos en riesgo (sin falla y no censurados) en t, respectivamente [25].

Cada individuo observado falla o está censurado en algún tiempo t. En el caso de que el individuo i falle en t_i, $dN_i(t_i) = 1$ y $Y_i(s) = I(s \leqslant t_i)$, luego

$$\prod_{t=0}^{\infty} h(t)^{dN_i(t)}[1-h(t)]^{Y_i(t)(1-dN_i(t))} = h(t_i)\prod_{t=0}^{t_i-1}[1-h(t)],$$

usando (1.12) y (1.10) se concluye que

$$\prod_{t=0}^{\infty} h(t)^{dN_i(t)}[1-h(t)]^{Y_i(t)(1-dN_i(t))} = h(t_i)S(t_{i-1}) = f(t_i).$$

Mientras que en el caso de que el individuo sea censurado en t_i, $dN_i(t_i) = 0$ y $Y_i(s) = I(s \leqslant t_i)$, por lo tanto

$$\prod_{t=0}^{\infty} h(t)^{dN_i(t)}[1-h(t)]^{Y_i(t)(1-dN_i(t))} = \prod_{t=0}^{t_i}[1-h(t)] = S(t_i).$$

Se concluye que en la notación (t_i, δ_i), (2.9) es igual a

$$L \propto \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}. \tag{2.11}$$

Distribuciones continuas

Para obtener la verosimilitud en el caso de distribuciones continuas se asocian $dN_i(t)$ y $dC_i(t)$ con un intervalo pequeño $[t, t+dt)$ en una partición del eje del tiempo, es decir

$$\begin{aligned} dN_i(t) &= Y_i(t)I(T_i \in [t, t+dt)) \\ dC_i(t) &= Y_i(t)I(\text{El individuo } i \text{ está censurado en } t \in [t, t+dt)), \end{aligned}$$

además, $\mathcal{H}(t) = \{(\underline{dN(s)}, \underline{dC(s)}) | s < t\}$. En (2.7) y (2.8) se reemplaza $h(t)$ por $h(t)dt$ de modo que

$$Pr(\underline{dN(t)}|\mathcal{H}(t)) = \prod_{i=1}^{n}(h(t)dt)^{dN_i(t)}[1-(h(t)dt)]^{Y_i(t)(1-dN_i(t))}$$

y

$$Pr(dN_i(t) = 1|\mathcal{H}(t)) = Y_i(t)(h(t)dt).$$

El argumento precedente se cumple esencialmente sin cambios cuando se toma la integral producto (Apéndice B) de (2.9), es decir

$$L \propto \prod_{i=1}^{n} \prod_{0}^{\infty} (h(t)dt)^{dN_i(t)}[1 - (h(t)dt)]^{Y_i(t)(1-dN_i(t))}. \tag{2.12}$$

Como en el caso discreto, cada individuo observado falla o está censurado en algún tiempo t. En el caso de que el individuo i falle en t_i, $dN_i(t_i) = 1$ y $Y_i(s) = I(s \leqslant t_i)$, usando (B.1) del Apéndice B, se tiene que

$$\begin{aligned}&\prod_{0}^{\infty} (h(t)dt)^{dN_i(t)}[1 - (h(t)dt)]^{Y_i(t)(1-dN_i(t))} \\ &= \prod_{0}^{t_i} (h(t)dt)^{dN_i(t)}[1 - (h(t)dt)]^{Y_i(t)(1-dN_i(t))}\end{aligned}$$

y

$$\prod_{t_i}^{\infty} (h(t)dt)^{dN_i(t)}[1 - (h(t)dt)]^{Y_i(t)(1-dN_i(t))} = h(t_i) \prod_{0}^{t_i} [1 - h(t)dt].$$

Dado que la distribución de T es continua, $H(\cdot)$ también es continua y $H'(t) = h(t)$. Luego de (B.2) se sigue que

$$\prod_{0}^{\infty} (h(t)dt)^{dN_i(t)}[1 - (h(t)dt)]^{Y_i(t)(1-dN_i(t))} = h(t_i) \exp\left\{-\int_0^{t_i} h(t)dt\right\},$$

utilizando (1.5) y luego (1.4) se tiene que

$$\prod_{0}^{\infty} (h(t)dt)^{dN_i(t)}[1 - (h(t)dt)]^{Y_i(t)(1-dN_i(t))} = h(t_i)S(t_i) = f(t_i).$$

Si el individuo i está censurado en t_i, $dN_i(t_i) = 0$ y $Y_i(s) = I(s \leqslant t_i)$, luego

$$\begin{aligned}&\prod_{0}^{\infty} (h(t)dt)^{dN_i(t)}[1 - (h(t)dt)]^{Y_i(t)(1-dN_i(t))} \\ &= \prod_{0}^{t_i} (h(t)dt)^{dN_i(t)}[1 - (h(t)dt)]^{Y_i(t)(1-dN_i(t))}\end{aligned}$$

y

$$\begin{aligned}\prod_{t_i}^{\infty}(h(t)dt)^{dN_i(t)}[1-(h(t)dt)]^{Y_i(t)(1-dN_i(t))} &= \prod_{0}^{t_i}[1-h(t)dt] \\ &= \exp\left\{-\int_0^{t_i} h(u)du\right\} \\ &= S(t_i).\end{aligned}$$

Se concluye que en la notación (t_i, δ_i), (2.12) es equivalente a

$$L \propto \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}, \tag{2.13}$$

que es igual a (2.11).

Hay dos características adicionales importantes en el desarrollo precedente. Una es que se permite que el mecanismo de censura en el tiempo t dependa de la historia de la censura y falla antes de t. El proceso de censura tipo II es de este tipo. Más generalmente, sería permitido en un estudio tomar una decisión sobre censura de individuos (i.e. quitarlos del estudio) o terminar el estudio en el tiempo t de acuerdo a la información de falla hasta tal tiempo. Un segundo punto es que (2.9) está disponible sin un modelo específico del proceso de censura. Mientras que los términos $Pr(\underline{dC(t)}|\underline{dN(t)}, \mathcal{H}(t))$ en (2.6) no involucren parámetros de interés, se pueden eliminar de la verosimilitud, este es generalmente el caso con censura independiente [25].

La expresión (2.9) tiene la misma forma para una variedad de mecanismos de censura por la derecha. Además, los procedimientos de inferencia basados en la teoría de máxima verosimilitud para muestras grandes se puede aplicar de manera directa. Sin embargo, las distribuciones de probabilidad en las cuales se basa (2.9) pueden diferir substancialmente de acuerdo a los procesos de censura y las propiedades de los estimadores o pruebas pueden ser diferentes en muestras pequeñas [25].

Ahora, suponga que se ha elegido una familia específica, así que se sabe que la distribución del tiempo de falla depende de un vector paramétrico $\underline{\phi}$ y que está disponible para la inferencia sobre $\underline{\phi}$ una sola muestra aleatoria de tiempos de falla, posiblemente sujeta a censura por la derecha. Frecuentemente se puede escribir $\underline{\phi}' = (\underline{\omega}', \underline{\lambda}')$, en donde $\underline{\omega}$ es un parámetro de interés particular y $\underline{\lambda}$ un parámetro de ruido. En este caso se tiene que la verosimilitud completa de n individuos independientes, indexados por i es

$$L(\underline{\phi}; \underline{t}) = \prod_{i \in \mathscr{U}} f(t_i; \underline{\phi}) \prod_{i \in \mathscr{C}} S(t_i; \underline{\phi}), \tag{2.14}$$

en donde $\mathscr{U}$ y $\mathscr{C}$ son los conjuntos de los individuos no censurados y censurados, respectivamente. Luego el $\log$ de la verosimilitud es

$$l(\underline{\phi};\underline{t}) = \log\left(L(\underline{\phi};\underline{t})\right) = \sum_{i\in\mathscr{U}} \log(f(t_i;\underline{\phi})) + \sum_{i\in\mathscr{C}} \log(S(t_i;\underline{\phi})). \tag{2.15}$$

Si la distribución del tiempo de falla es continua, se tiene que

$$f(t) = h(t)S(t)$$

por lo que se puede escribir

$$l(\underline{\phi};\underline{t}) = \sum_{i\in\mathscr{U}} \log(h(t_i;\underline{\phi})) + \sum_{i\in\mathscr{U}\cup\mathscr{C}} \log(S(t_i;\underline{\phi})).$$

Como el $\log$ de la función de supervivencia es menos la función de riesgo acumulado, se tiene que

$$l(\underline{\phi};\underline{t}) = \sum_{i\in\mathscr{U}} \log(h(t_i;\underline{\phi})) - \sum_{i\in\mathscr{U}\cup\mathscr{C}} H(t_i;\underline{\phi}).$$

Finalmente, sea $r(u) = card\{i : t_i \geqslant u\}$, el número de sujetos que se consideran en el tiempo u, se puede representar a $l(\cdot;\underline{t})$ como

$$l(\underline{\phi};\underline{t}) = \sum_{i\in\mathscr{U}} \log(h(t_i;\underline{\phi})) - \int_0^\infty r(u)h(u;\underline{\phi})du. \tag{2.16}$$

La integral es formalmente sobre un rango infinito, ya que $r(u)$ es cero después del último tiempo de falla o tiempo censurado observado. El integrando puede ser interpretado como el riesgo total operando en el tiempo u [9].

2.2. La función de verosimilitud de datos censurados por intervalo

Ahora, se hallará la función de verosimilitud en el caso de datos censurados por intervalo, para esto suponga que en un estudio los individuos se observan intermitentemente en tiempos discretos. Considérese que el individuo i, $i = 1, 2, \ldots, n$, se observa en un conjunto de tiempos preespecificados $0 = t_{i0} < t_{i1} < \ldots < t_{im_i} < \infty$. Si un individuo no ha fallado en el tiempo $t_{i,j-1}$ ($j = 1, \ldots, m_i$), se observa en el siguiente tiempo, t_{ij}, y se determina si ocurrió la falla o no en el intervalo $(t_{i,j-1}, t_{ij}]$. Entonces los datos observados consisten de un intervalo $(U_i, V_i]$ para cada individuo, es decir se tiene

la información de que $U_i < T_i \leqslant V_i$ y el tiempo de falla se dice censurado por intervalo. Si la falla no ocurrió para el tiempo t_{im_i}, entonces $V_i = \infty$ y $U_i = t_{im_i}$ es un tiempo censurado por la derecha para T_i.

Dado que la observación para el i-ésimo individuo es multinomial con probabilidades $p_{ij} = Pr(t_{i,j-1} < T_i \leqslant t_{ij}) = F(t_{ij}) - F(t_{i,j-1})$, $j = 1, \ldots, m_i$ y $p_{i,m_i+1} = Pr(t_{i,m_i} < T_i) = 1 - F(t_{i,m_i})$, la función de verosimilitud observada de una muestra de n individuos independientes con censura por intervalo es

$$L = \prod_{i=1}^{n} [F(V_i) - F(U_i)], \tag{2.17}$$

donde $F(\cdot)$ es la f.d.a. para T.

Para modelos paramétricos, la inferencia basada en la verosimilitud (2.17) cae bajo la metodología de máxima verosimilitud.

La suposición de que los tiempos de observación t_{ij} se fijan de antemano, o incluso que se determinan independientemente del proceso que genera tiempos de falla, no es adecuado en muchos entornos. Por ejemplo, una decisión respecto a cuando se verá a un individuo en un estudio clínico se puede basar en la información actual sobre el individuo. Más generalmente, suponga que si un individuo no ha fallado y no está censurado en el tiempo $t_{i,j-1}$, se decide sobre el siguiente tiempo de observación t_{ij} basándose en las fallas observadas e historia del tiempo de observación hasta $t_{i,j-1}$, $\mathcal{H}(t_{i,j-1})$ (que incluye la información de que el individuo i no ha fallado y no está censurado en el tiempo $t_{i,j-1}$). Sin embargo, la elección de t_{ij} es condicionalmente independiente de la información de fallas más allá de $t_{i,j-1}$, dado $\mathcal{H}(t_{i,j-1})$ [25].

Bajo este proceso de observación los datos consisten de los tiempos de observación $0 = t_{i0} < t_{i1} < \ldots < t_{im_i} \leqslant \infty$ y la información de que $t_{i,m_i-1} < T_i \leqslant t_{im_i}$. Notar que $t_{im_i} = \infty$ corresponde al tiempo de falla censurado por la derecha en el tiempo t_{i,m_i-1}. Asumiendo que los términos $Pr(t_{ij}|\mathcal{H}(t_{i,j-1}))$, $j = 1, \ldots, m_i$ no contienen información sobre la distribución de tiempo de falla, la función de verosimilitud observada para el individuo i es proporcional a

$$\left\{ \prod_{j=1}^{m_i-1} Pr(T_i > t_{ij}|\mathcal{H}(t_{i,j-1}), t_{ij}) \right\} Pr(T_i \leqslant t_{im_i}|\mathcal{H}(t_{i,m_i-1}), t_{im_i}). \tag{2.18}$$

Como la elección de t_{ij} es condicionalmente independiente de la informa-

ción de fallas más allá de $t_{i,j-1}$, dado $\mathcal{H}(t_{i,j-1})$ se sigue que

$$\begin{aligned} Pr(T_i > t_{ij}|\mathcal{H}(t_{i,j-1}), t_{ij}) &= Pr(T_i > t_{ij}|\mathcal{H}(t_{i,j-1})) \\ &= \frac{1 - F(t_{ij})}{1 - F(t_{i,j-1})} \end{aligned}$$

y

$$\begin{aligned} Pr(T_i \leqslant t_{im_i}|\mathcal{H}(t_{i,m_i-1}), t_{im_i}) &= 1 - Pr(T_i > t_{im_i}|\mathcal{H}(t_{i,m_i-1}), t_{im_i}) \\ &= 1 - \frac{1 - F(t_{im_i})}{1 - F(t_{i,m_i-1})} \\ &= \frac{F(t_{im_i}) - F(t_{i,m_i-1})}{1 - F(t_{i,m_i-1})}, \end{aligned}$$

en donde $F(\cdot)$ es la f.d.a. para T. Se sigue que (2.18) es igual a

$$\left\{ \prod_{j=1}^{m_i-1} \frac{1 - F(t_{ij})}{1 - F(t_{i,j-1})} \right\} \frac{F(t_{im_i}) - F(t_{i,m_i-1})}{1 - F(t_{i,m_i-1})} = F(t_{im_i}) - F(t_{i,m_i-1}).$$

Por lo tanto la verosimilitud es

$$L \propto \prod_{i=1}^{n} [F(t_{im_i}) - F(t_{i,m_i-1})],$$

que es de la misma forma de (2.17).

Como en el caso de censura por la derecha, suponga que se ha elegido una familia específica, así que se sabe que la distribución del tiempo de falla depende de un vector paramétrico $\underline{\phi}$ y que está disponible para la inferencia sobre $\underline{\phi}$ una sola muestra de tiempos de falla, sujeta a censura por intervalo. En este caso se tiene que la verosimilitud completa de n individuos independientes, indexados por i es

$$L(\underline{\phi}; t) = \prod_{i=1}^{n} [F(t_{im_i}; \underline{\phi}) - F(t_{i,m_i-1}; \underline{\phi})]$$

y el $\log$ de la verosimilitud es

$$l(\underline{\phi}; t) = \log\left(L(\underline{\phi}; t)\right) = \sum_{i=1}^{n} \log[F(t_{im_i}; \underline{\phi}) - F(t_{i,m_i-1}; \underline{\phi})].$$

2.3. Inferencia basada en la función de verosimilitud

Cuando se ha elegido una familia específica y la distribución del tiempo de falla depende del vector paramétrico $\underline{\phi}' = (\underline{\omega}', \underline{\lambda}')$, se puede llevar a cabo un procedimiento asintótico, basado en la función de verosimilitud, para contrastar las hipótesis $H_0 : \underline{\omega} = \underline{\omega}_0$ contra $H_0 : \underline{\omega} \neq \underline{\omega}_0$. El estadístico de razón de verosimilitud es

$$W(t) = 2[l(\widehat{\underline{\omega}}, \widehat{\underline{\lambda}}; t) - l(\underline{\omega}_0, \widehat{\underline{\lambda}}_{\underline{\omega}_0}; t)],$$

en donde $(\widehat{\underline{\omega}}', \widehat{\underline{\lambda}}')$ es el estimador de máxima verosimilitud de $\underline{\phi}' = (\underline{\omega}', \underline{\lambda}')$ y $\widehat{\underline{\lambda}}_{\underline{\omega}_0}$ es el estimador de máxima verosimilitud de $\underline{\lambda}$ cuando $\underline{\omega} = \underline{\omega}_0$. Bajo la hipótesis nula, $W(t)$ es asintóticamente distribuida como una variable aleatoria χ^2 con p grados de libertad, en donde p es igual a la diferencia entre el número de parámetros independientes en el espacio paramétrico y el número de parámetros independientes en la hipótesis nula. La prueba de razón de verosimilitud de tamaño α para contrastar H_0 contra H_1 tiene región de rechazo

$$\{t : W(t) > c_{p,\alpha}\},$$

en donde $c_{p,\alpha}$ es el $1 - \alpha$ cuantil de una distribución $\chi^2_{(p)}$.

2.4. Algunas distribuciones de tiempo de falla

Varias familias paramétricas se usan como modelos en el análisis de datos de tiempo de fallas. Entre los modelos univariados, algunas distribuciones ocupan una posición central ya que se ha demostrado su utilidad en una amplia gama de situaciones. Principalmente en esta categoría están la distribuciones exponencial, Weibull, $\log$ normal, $\log$ logística y gama [25]. A continuación se presentan estas distribuciones agregando la distribución gama generalizada y algunas de sus propiedades.

2.4.1. Distribución exponencial

Históricamente, el primer modelo de distribución de tiempo de falla ampliamente discutido fue el exponencial. Esto se debió en parte a la disponibilidad de métodos estadísticos simples para este [25]. Si T tiene una distribución exponencial con parámetro λ positivo ($T \sim Exp(\lambda)$), entonces

$$\begin{aligned} f(t) &= \lambda e^{-\lambda t}, \text{ para } t \geqslant 0, \\ F(t) &= 1 - e^{-\lambda t}, \text{ para } t \geqslant 0, \end{aligned}$$

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

y

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

La distribución donde $\lambda = 1$ se llama distribución exponencial estándar.

Luego de (1.1), (1.4) y (1.6) se sigue que

$$S(t) = e^{-\lambda t}, \text{ para } t \geqslant 0,$$

$$h(t) = \lambda, \text{ para } t \geqslant 0$$

y

$$H(t) = \lambda t, \text{ para } t \geqslant 0.$$

La proporción de falla instantánea es independiente de t, así que la probabilidad condicional de falla en un intervalo de tiempo de longitud especificada es la misma independientemente de cuánto tiempo el individuo ha estado en el estudio; esto se refiere como propiedad de pérdida de memoria de la distribución exponencial [19]. La suposición de la función de riesgo constante es muy restrictiva, por lo que las aplicaciones del modelo son bastante limitadas [25].

Una revisión empírica de la distribución exponencial para un conjunto de datos de supervivencia es proporcionada por la gráfica del $\log$ de la función de supervivencia estimada contra t. Tal gráfica debe aproximarse a una línea recta que pasa por el origen [19].

El p-ésimo cuantil de la distribución exponencial es el valor t_p que cumple que $F(t_p) = p$, esto es, $t_p = -\frac{1}{\lambda}\log(1-p)$.

Una gran variedad de mecanismos de censura por la derecha cumplen que: si se supone que los tiempos de falla de una muestra de tamaño n, (t_i, δ_i), $i = 1, \ldots, n$, sujeta a censura por la derecha tienen una distribución $Exp(\lambda)$, la función de verosimilitud observada para el parámetro del tiempo de falla está dada por

$$L(\lambda) = \lambda^d e^{-\lambda \sum_{i=1}^{n} t_i},$$

en donde d es el número total de fallas observadas. Luego la función $\log$ de verosimilitud está dada por

$$l(\lambda) = \log(L(\lambda)) = d\log(\lambda) - \lambda \sum_{i=1}^{n} t_i.$$

La primera derivada de $l(\cdot)$ es

$$\frac{\partial l}{\partial \lambda}(\lambda) = \frac{d}{\lambda} - \sum_{i=1}^{n} t_i,$$

por lo que el único punto crítico de $l(\cdot)$ es $\lambda = \frac{d}{\sum_{i=1}^{n} t_i}$. Como $\frac{\partial^2 l}{\partial^2 \lambda}(\lambda) = -\frac{d}{\lambda^2}$ se tiene que

$$\frac{\partial^2 l}{\partial^2 \lambda}\left(\frac{d}{\sum_{i=1}^{n} t_i}\right) = -\frac{d}{\left(\frac{d}{\sum_{i=1}^{n} t_i}\right)^2} < 0,$$

por lo que se concluye que el estimador de máxima verosimilitud para λ es

$$\widehat{\lambda} = \frac{d}{\sum_{i=1}^{n} t_i}.$$

2.4.2. Distribución Weibull

La distribución Weibull es quizás el modelo de distribución de falla más ampliamente utilizado. Es común la aplicación a los tiempos de vida o la durabilidad de elementos manufacturados, se usa como modelo en diversos tipos de elementos tales como componentes del automóvil y aislamiento eléctrico. También se usa en aplicaciones biológicas y médicas, por ejemplo, en estudios sobre el tiempo para la ocurrencia de tumores en poblaciones humanas o en animales de laboratorio [25].

Si T tiene una distribución Weibull con parámetro de escala λ y parámetro de forma θ, ambos positivos ($T \sim W(\lambda, \theta)$), entonces

$$\begin{aligned} f(t) &= \lambda\theta(\lambda t)^{\theta-1}\exp\{-(\lambda t)^{\theta}\}, \text{ para } t \geqslant 0, \\ F(t) &= 1 - \exp\{-(\lambda t)^{\theta}\}, \text{ para } t \geqslant 0, \end{aligned}$$

$$E(T) = \frac{\Gamma(1+1/\theta)}{\lambda}$$

y

$$Var(T) = \frac{\Gamma(1+2/\theta) - (\Gamma(1+1/\theta))^2}{\lambda^2},$$

en donde

$$\Gamma(k) = \int_0^{\infty} u^{k-1}e^{-u}du, \text{ para } k > 0,$$

es la función gama. Esta incluye a la distribución exponencial como el caso especial en donde $\theta = 1$. Luego de (1.1), (1.4) y (1.6) se sigue que

$$S(t) = \exp\{-(\lambda t)^{\theta}\}, \text{ para } t \geqslant 0,$$
$$h(t) = \lambda\theta(\lambda t)^{\theta-1}, \text{ para } t \geqslant 0$$

y

$$H(t) = (\lambda t)^{\theta}, \text{ para } t \geqslant 0.$$

La función de riesgo de la distribución Weibull es monótona creciente si $\theta > 1$, monótona decreciente si $\theta < 1$ y constante para $\theta = 1$ (éste es el caso de la distribución exponencial). El modelo es bastante flexible y se ha encontrado que proporciona una buena descripción de muchos tipos de datos de tiempos de falla. Esto y el hecho de que el modelo tiene expresiones simples para la f.d.p., f.d.a., función de supervivencia y función de riesgo es la razón de su popularidad [25].

Dado que $\log(-\log(S(t))) = \theta(\log\lambda + \log t)$, una revisión empírica para la distribución Weibull es proporcionada por una gráfica de $\log(-\log(\widehat{S}(t)))$ contra $\log t$, en donde $\widehat{S}(\cdot)$ es un estimador muestral de la función de supervivencia (Kaplan-Meier). La gráfica debe dar aproximadamente una línea recta, la pendiente y la intersección con el eje horizontal de la recta proporcionan una estimación aproximada de θ y $-\log\theta$, respectivamente [19].

El p-ésimo cuantil de la distribución de Weibull es el valor t_p que cumple que $F(t_p) = p$, esto es, $t_p = \dfrac{1}{\lambda}(-\log(1-p))^{1/\theta}$.

Una gran variedad de mecanismos de censura por la derecha cumplen que: si se supone que los tiempos de falla de una muestra de tamaño n, (t_i, δ_i), $i = 1, \ldots, n$, sujeta a censura por la derecha tienen una distribución $W(\lambda, \theta)$, de (2.14) la función de verosimilitud observada para los parámetros del tiempo de falla está dada por

$$L(\lambda, \theta) = \lambda^{d\theta}\theta^{d}e^{-\lambda^{\theta}\sum_{i=1}^{n} t_i^{\theta}} \prod_{i\in\mathscr{U}} t_i^{\theta-1},$$

en donde d es el número total de fallas observadas y $\mathscr{U}$ es el conjunto de individuos no censurados. Luego la función $\log$ de verosimilitud está dada por

$$l(\lambda, \theta) = \log(L(\lambda, \theta)) = d\log(\lambda^{\theta}\theta) + (\theta - 1)\sum_{i\in\mathscr{U}} \log(t_i) - \lambda^{\theta}\sum_{i=1}^{n} t_i^{\theta}. \quad (2.19)$$

Las primeras derivadas de $l(\cdot)$ son

$$\frac{\partial l}{\partial\lambda}(\lambda, \theta) = \frac{d\theta}{\lambda} - \theta\lambda^{\theta-1}\sum_{i=1}^{n} t_i^{\theta}$$

y

$$\frac{\partial l}{\partial \theta}(\lambda,\theta) = \frac{d}{\theta} + d\log\lambda + \sum_{i\in\mathscr{U}} \log t_i - \lambda^{\theta}\sum_{i=1}^{n} t_i^{\theta}\log(\lambda t_i). \tag{2.20}$$

Para θ fijo, el estimador de máxima verosimilitud para λ se obtiene resolviendo $\dfrac{\partial l}{\partial \lambda}(\lambda,\theta) = 0$, de donde

$$\widehat{\lambda} = \left(\frac{d}{\sum_{i=1}^{n} t_i^{\theta}}\right)^{1/\theta}.$$

Mientras que el estimador de máxima verosimilitud para θ se obtiene resolviendo

$$\frac{\partial l}{\partial \theta}(\widehat{\lambda},\theta) = \frac{d}{\theta} + \sum_{i\in\mathscr{U}} \log(t_i) - \frac{d\sum_{i=1}^{n} t_i^{\theta}\log(t_i)}{\sum_{i=1}^{n} t_i^{\theta}} = 0,$$

lo cual no se puede hacer analíticamente, por lo que se usan métodos numéricos como el de Newton-Raphson, el de la regla falsa, el de bisección, etc.

2.4.3. Distribución $\log$ normal

La distribución $\log$ normal se ha usado como modelo en diversas aplicaciones en ingeniería, medicina y otras áreas [25]. Se dice que el tiempo de falla T tiene una distribución $\log$ normal con parámetros μ y σ^2 ($T \sim \log N(\mu,\sigma^2)$), si $\log(T)$ se distribuye normalmente con media μ y varianza σ^2 ($\log(T) \sim N(\mu,\sigma^2)$), $\mu \in \mathbb{R}$ y σ^2 positivo. En este caso se tiene que

$$f(t) = \frac{1}{(2\pi\sigma^2)^{1/2}t}\exp\left\{-\frac{1}{2}\left(\frac{\log t-\mu}{\sigma}\right)^2\right\}, \text{ para } t>0$$

y

$$\begin{aligned} F(t) &= \int_0^t \frac{1}{(2\pi\sigma^2)^{1/2}x}\exp\left\{-\frac{1}{2}\left(\frac{\log x-\mu}{\sigma}\right)^2\right\}dx \\ &= \int_{-\infty}^{\frac{\log t-\mu}{\sigma}} \frac{1}{(2\pi)^{1/2}}\exp\left\{-\frac{1}{2}z^2\right\}dz \\ &= \Phi\left(\frac{\log t-\mu}{\sigma}\right), \text{ para } t \geqslant 0, \end{aligned}$$

en donde

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

es la función de distribución normal estándar. Además

$$E(T) = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

y

$$Var(T) = \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2).$$

Luego, de (1.1) y (1.4) se sigue que

$$S(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right), \text{ para } t \geqslant 0$$

y

$$h(t) = \frac{\exp\left\{-\frac{1}{2}\left(\frac{\log t - \mu}{\sigma}\right)^2\right\}}{(2\pi\sigma^2)^{1/2} t \left(1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right)\right)}, \text{ para } t \geqslant 0.$$

La función de riesgo cumple que $\lim_{t\to 0+} h(t) = 0$, crece hasta alcanzar el máximo y después decrece aproximándose a 0 cuanto $t \to \infty$. La mediana de la distribución $\log$ normal es $t_{0.5} = \exp(\mu)$ [25].

2.4.4. Distribución $\log$ logística

Si T tiene una distribución $\log$ logística con parámetros α y θ, ambos positivos ($T \sim LLogist(\alpha, \theta)$), entonces

$$\begin{aligned} f(t) &= \frac{(\theta/\alpha)(t/\alpha)^{\theta-1}}{(1 + (t/\alpha)^{\theta})^2}, \text{ para } t \geqslant 0, \\ F(t) &= 1 - (1 + (t/\alpha)^{\theta})^{-1}, \text{ para } t \geqslant 0 \end{aligned}$$

y

$$E(T^r) = \alpha^r \Gamma(1 + r/\theta)\Gamma(1 - r/\theta), \text{ si } \theta > r.$$

Luego, de (1.1) y (1.4) se sigue que

$$S(t) = (1 + (t/\alpha)^{\theta})^{-1}, \text{ para } t \geqslant 0$$

y

$$h(t) = \frac{(\theta/\alpha)(t/\alpha)^{\theta-1}}{(1 + (t/\alpha)^{\theta})}, \text{ para } t \geqslant 0.$$

La distribución log logística tiene su nombre debido a que $\log(T)$ tiene una distribución logística con parámetros $u = \log(\alpha) \in \mathbb{R}$ y $b = \theta^{-1}$ positivo ($\log(T) \sim Logist(u, b)$).

Para $\theta > 1$ la función de riesgo tiene la misma forma característica que la log normal, esto es $\lim_{t\to 0+} h(t) = 0$, crece hasta un máximo y después decrece y se aproxima a 0 monótonamente cuanto $t \to \infty$. Para $\theta \leqslant 1$ la función de riesgo es monótona decreciente [25].

2.4.5. Distribución gama

Si T tiene una distribución gama con parámetro de escala λ^{-1} y parámetro de forma k, ambos positivos ($T \sim G(k, \lambda)$), entonces

$$f(t) = \frac{\lambda(\lambda t)^{k-1}e^{-\lambda t}}{\Gamma(k)}, \text{ para } t \geqslant 0$$

y

$$\begin{aligned} F(t) &= \frac{1}{\Gamma(k)} \int_0^t \lambda(\lambda x)^{k-1}e^{-\lambda x}dx \\ &= \frac{1}{\Gamma(k)} \int_0^{\lambda t} u^{k-1}e^{-u}du \\ &= I(k, \lambda t), \text{ para } t \geqslant 0, \end{aligned}$$

en donde

$$I(k, x) = \frac{1}{\Gamma(k)} \int_0^x u^{k-1}e^{-u}du$$

es la función gama incompleta [25]. Además

$$E(T) = k/\lambda$$

y

$$Var(T) = k/\lambda^2.$$

Luego de (1.1) y (1.4) se sigue que

$$S(t) = 1 - I(k, \lambda t), \text{ para } t \geqslant 0$$

y

$$h(t) = \frac{\lambda(\lambda t)^{k-1}e^{-\lambda t}}{\Gamma(k)(1 - I(k, \lambda t))}, \text{ para } t \geqslant 0.$$

La función de riesgo es monótona creciente para $k > 1$ con $h(0) = 0$ y $\lim_{t\to\infty} h(t) = \lambda$. Para $0 < k < 1$, $h(\cdot)$ es monótona decreciente con $\lim_{t\to 0^+} h(t) = \infty$ y $\lim_{t\to\infty} h(t) = \lambda$ [25].

Cuando $k = 1$ esta distribución es la exponencial. La distribución gama con $\lambda = 1$ es llamada la distribución gama de un parámetro o gama estándar ($T \sim G(k)$) [25]. Con k entero, la distribución gama algunas veces es llamada una distribución de Erlang especial [19].

La distribución gama no se utiliza tanto como modelo de tiempo de falla como las distribuciones Weibull, $\log$ normal y $\log$ logística. Sin embargo, ésta ajusta adecuadamente a una variedad de datos de tiempo de falla y también surge en algunas aplicaciones que involucran la distribución exponencial, ya que la suma de variables aleatorias exponenciales independientes e idénticamente distribuidas (i.i.d.) tiene una distribución gama [25].

2.4.6. Distribución gama generalizada

En el Análisis de Supervivencia, distribuciones tales como la gama, Weibull, exponencial y $\log$ normal juegan un papel muy importante, todas ellas pertenecen a la familia de la distribución gama generalizada (GG) para valores específicos de los parámetros [18].

Si T tiene una distribución GG con parámetros de forma $\theta > 0$, $k > 0$ y de escala $\alpha > 0$ ($T \sim GG(\alpha, k, \theta)$), entonces

$$f(t) = \frac{\theta}{\alpha\Gamma(k)} \left(\frac{t}{\alpha}\right)^{\theta k-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\theta}\right\}, \text{ para } t \geqslant 0$$

y

$$\begin{aligned} F(t) &= \frac{1}{\Gamma(k)} \int_0^t \frac{\theta}{\alpha} \left(\frac{s}{\alpha}\right)^{\theta k-1} \exp\left\{-\left(\frac{s}{\alpha}\right)^{\theta}\right\} ds \\ &= \frac{1}{\Gamma(k)} \int_0^{\left(\frac{t}{\alpha}\right)^{\theta}} u^{k-1} \exp(-u) du \\ &= I\left(k, \left(\frac{t}{\alpha}\right)^{\theta}\right), \text{ para } t \geqslant 0. \end{aligned}$$

Además

$$E(T) = \frac{\alpha\Gamma\left(k + \frac{1}{\theta}\right)}{\Gamma(k)}$$

y

$$Var(T) = \frac{\alpha^2}{\Gamma^2(k)}\left[\Gamma(k)\Gamma\left(k+\frac{2}{\theta}\right) - \Gamma^2\left(k+\frac{1}{\theta}\right)\right].$$

Luego de (1.1) y (1.4) se sigue que

$$S(t) = 1 - I\left(k, \left(\frac{t}{\alpha}\right)^{\theta}\right), \text{ para } t \geqslant 0$$

y

$$h(t) = \frac{\theta\left(\frac{t}{\alpha}\right)^{\theta k-1}\exp\left\{-\left(\frac{t}{\alpha}\right)^{\theta}\right\}}{\alpha\Gamma(k)\left\{1 - I\left(k, \left(\frac{t}{\alpha}\right)^{\theta}\right)\right\}}, \text{ para } t \geqslant 0.$$

Las subfamilias de la GG son la exponencial cuando $k = \theta = 1$, la gama para $\theta = 1$ y la Weibull si $k = 1$. La distribución $\log$ normal se obtiene como una distribución límite cuando $k \to \infty$. Cuando $\theta = 2$ se obtiene una subfamilia de la GG que es conocida como la distribución normal generalizada [21].

Aunque resulta natural pensar en trabajar con la distribución GG en vez de cada una de las distribuciones consideradas anteriormente, hay un inconveniente, y es que estimar los parámetros de la distribución GG no es tan sencillo, por ello muchos estadísticos prefieren trabajar por separado cada distribución [18].

En la Tabla 2.1 se muestran las distribuciones de tiempo de falla presentadas en esta sección con algunas de las características que las especifican.

2.5. Modelos no paramétricos

Ahora se discuten técnicas no paramétricas, i.e. que no requieren especificaciones de la forma funcional de la distribución que los tiempos de falla tendrían en la ausencia de censura.

2.5.1. Estimador Producto-Límite (Estimador Kaplan-Meier)

Supóngase que la posible distribución impropia es discreta con probabilidad $f(t_j)$ en los tiempos especificados $t_1 < t_2 < \ldots < t_g$. En la práctica, estos puntos se toman a menudo igualmente espaciados, $t_j = j$ en unidades de tiempo adecuadas, pero esto no es necesario.

	$T\sim$	$f(t)$, para $t\geqslant 0$	$F(t)$, para $t\geqslant 0$	$S(t)$, para $t\geqslant 0$	$h(t)$, para $t\geqslant 0$	$E(T)$	$Var(T)$
Exponencial	$Exp(\lambda)$ $\lambda>0$	$\lambda e^{-\lambda t}$	$1-e^{-\lambda t}$	$e^{-\lambda t}$	λ	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Weibull	$W(\lambda,\theta)$ $\lambda,\theta>0$	$\lambda\theta(\lambda t)^{\theta-1}\exp\{-(\lambda t)^{\theta}\}$	$1-\exp\{-(\lambda t)^{\theta}\}$	$\exp\{-(\lambda t)^{\theta}\}$	$\lambda\theta(\lambda t)^{\theta-1}$	$\frac{\Gamma(1+1/\theta)}{\lambda}$	$\frac{\Gamma(1+2/\theta)-(\Gamma(1+1/\theta))^2}{\lambda^2}$
log normal	$\log N(\mu,\sigma^2)$ $\mu\in\mathbb{R}$, $\sigma^2>0$	$\frac{1}{(2\pi\sigma^2)^{1/2}t}\exp\left\{-\frac{1}{2}\left(\frac{\log t-\mu}{\sigma}\right)^2\right\}$	$\Phi\left(\frac{\log t-\mu}{\sigma}\right)$	$1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)$	$\frac{\exp\left\{-\frac{1}{2}\left(\frac{\log t-\mu}{\sigma}\right)^2\right\}}{(2\pi\sigma^2)^{1/2}t\left(1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)\right)}$	$\exp\left(\mu+\frac{\sigma^2}{2}\right)$	$\exp(2\mu+2\sigma^2)-\exp(2\mu+\sigma^2)$
log logística	$LLogist(\alpha,\theta)$ $\alpha,\theta>0$	$\frac{(\theta/\alpha)(t/\alpha)^{\theta-1}}{(1+(t/\alpha)^{\theta})^2}$	$1-(1+(t/\alpha)^{\theta})^{-1}$	$(1+(t/\alpha)^{\theta})^{-1}$	$\frac{(\theta/\alpha)(t/\alpha)^{\theta-1}}{(1+(t/\alpha)^{\theta})}$	$\alpha\Gamma(1+\theta^{-1})\Gamma(1-\theta^{-1})$ si $\theta>1$	
Gama	$G(k,\lambda)$ $k,\lambda>0$	$\frac{\lambda(\lambda t)^{k-1}e^{-\lambda t}}{\Gamma(k)}$	$I(k,\lambda t)$	$1-I(k,\lambda t)$	$\frac{\lambda(\lambda t)^{k-1}e^{-\lambda t}}{\Gamma(k)(1-I(k,\lambda t))}$	$\frac{k}{\lambda}$	$\frac{k}{\lambda^2}$
Gama generalizada	$GG(\alpha,k,\theta)$ $\alpha,k,\theta>0$	$\frac{\theta}{\alpha\Gamma(k)}\left(\frac{t}{\alpha}\right)^{\theta k-1}\exp\left\{-\left(\frac{t}{\alpha}\right)^{\theta}\right\}$	$I\left(k,\left(\frac{t}{\alpha}\right)^{\theta}\right)$	$1-I\left(k,\left(\frac{t}{\alpha}\right)^{\theta}\right)$	$\frac{\theta\left(\frac{t}{\alpha}\right)^{\theta k-1}\exp\left\{-\left(\frac{t}{\alpha}\right)^{\theta}\right\}}{\alpha\Gamma(k)\left\{1-I\left(k,\left(\frac{t}{\alpha}\right)^{\theta}\right)\right\}}$	$\frac{\alpha\Gamma\left(k+\frac{1}{\theta}\right)}{\Gamma(k)}$	$\frac{\alpha^2}{\Gamma^2(k)}\left[\Gamma(k)\Gamma\left(k+\frac{2}{\theta}\right)-\Gamma^2\left(k+\frac{1}{\theta}\right)\right]$

Tabla 2.1: Distribuciones de tiempo de falla con algunas característica que las especifican: f.d.p. ($f(\cdot)$), f.d.a. ($F(\cdot)$), función de supervivencia ($S(\cdot)$), función de riesgo ($h(\cdot)$), media ($E(T)$) y Varianza ($Var(T)$).

De (1.12) se tiene que la función de supervivencia $S(\cdot)$ se puede expresar en términos de la función de riesgo discreta $h(\cdot)$ como

$$S(t) = \prod_{j:t_j \leqslant t} (1 - h(t_j)).$$

Luego, de (1.10) los $f(t_j)$ pueden ser escritos en términos de los $h(t_j)$ en la forma

$$f(t_j) = S(t_{j-1})h(t_j) = \left\{ \prod_{i:t_i < t_j} (1 - h(t_i)) \right\} h(t_j), j = 1, 2, \ldots.$$

Las restricciones de que $f(t_j) \geqslant 0$ y que $\sum f(t_j) \leqslant 1$ se convierten en, simplemente, $0 \leqslant h(t_j) \leqslant 1$.

Un estimador no paramétrico de la función de supervivencia es

$$\widehat{S}(t) = \prod_{j:t_j \leqslant t} (1 - \widehat{h}(t_j)), \tag{2.21}$$

en donde $\widehat{h}(t_j)$ es el estimador de máxima verosimilitud de $h(t_j)$. De la ecuación (2.10), el $\log$ de la función de verosimilitud de datos con censura por la derecha en términos de los $h(t_j)$ cumple que

$$l((h(t_1), h(t_2), \ldots, h(t_g))'; \underline{t}) = \sum_{j=1}^{g} \{d_j \log(h(t_j)) + (r_j - d_j) \log(1 - h(t_j))\},$$

en donde r_j es el número de individuos en riesgo en t_j y d_j es el número de tiempos de falla observados igual a t_j. Es una convención incluir en r_j cualquier individuo que esté censurado en t_j.

Luego, $\widehat{h}(t_j)$ es la solución de

$$\frac{\partial l}{\partial h(t_j)}((h(t_1), h(t_2), \ldots, h(t_g))'; \underline{t}) = 0 \Leftrightarrow \frac{d_j}{h(t_j)} - \frac{r_j - d_j}{1 - h(t_j)} = 0,$$

i.e. $\widehat{h}(t_j) = \dfrac{d_j}{r_j}$. El correspondiente estimador $\widehat{S}(\cdot)$ de la función de supervivencia, $S(\cdot)$, es

$$\widehat{S}(t) = \prod_{j:t_j \leqslant t} \left(1 - \frac{d_j}{r_j}\right), \text{ para } t \geqslant 0, \tag{2.22}$$

obtenida de la sustitución en (2.21).

Cualquier término en el producto que tenga $d_j = 0$ puede ser omitido sin afectar (2.22). El estimador $\widehat{S}(\cdot)$ es, por lo tanto, formalmente independiente de la selección de los puntos de t_j para los cuales el número de fallas observadas es cero. Así, $\widehat{S}(\cdot)$ es una función que sólo depende de los datos. Usualmente, $\widehat{S}(\cdot)$ es llamado el estimador de Kaplan-Meier o producto-Límite.

Si en una muestra de tamaño n no hay censura y los tiempos de falla son $t_1, \ldots, t_k$, entonces $r_1 = n$, $r_j = r_{j-1} - d_{j-1}$, $j = 2, \ldots, k$ y (2.22) se reduce a la función de supervivencia empírica,

$$\widehat{S}(t) = \frac{\text{Número de observaciones } > t}{n}, \text{ para } t \geqslant 0.$$

En ambos casos de censura y no censura, $\widehat{S}(\cdot)$ es una función escalonada continua por la derecha con $\widehat{S}(0) = 1$ y decrece por un factor $(r_j - d_j)/r_j$ en cada tiempo de falla t_j. El estimador no cambia en tiempos censurados, el efecto de los tiempos censurados esta reflejado en los valores r_j y así en el tamaño de los saltos de $\widehat{S}(\cdot)$.

Las distribuciones de tiempo continuas se pueden tratar como un límite del caso de tiempo discreto. Ahora, se piensa en la función de riesgo acumulado $H(\cdot)$ como el parámetro a estimar. Con $dH(t)$ como el incremento de la función de riesgo acumulado sobre $[t, t + dt)$, la verosimilitud (2.12) es igual a la integral producto

$$L \propto \prod_0^\infty dH(t)^{dN(t)}[1 - dH(t)]^{Y(t)-dN(t)}, \tag{2.23}$$

en donde $dN(t) = \sum_{i=1}^n dN_i(t)$ y $Y(t) = \sum_{i=1}^n Y_i(t)$. Si se considera a (2.23) con respecto al espacio de todas las funciones de riesgo acumulativo, $H(\cdot)$, ésta es maximizada por una función con saltos en cada uno de los tiempos de falla observados. Por comparación directa con el caso de tiempo discreto se tiene que L es maximizada por la función $\widehat{H}(\cdot)$ con incrementos

$$d\widehat{H}(t) = \frac{dN(t)}{Y(t)}, \text{ para } t \geqslant 0, Y(t) > 0. \tag{2.24}$$

Cuando $Y(t) = 0$, $d\widehat{H}(t)$ no está definido.

Luego por (1.5) y (B.2) se tiene que

$$\widehat{S}(t) = \exp\left(-\int_0^t \widehat{h}(u)du\right) = \prod_0^t (1 - d\widehat{H}(u)),$$

que es precisamente (2.22) [25].

Fórmula de Greenwood

Cuando se usa el estimador Producto-Límite es deseable tener un estimador de la varianza de $\widehat{S}(t)$. Siguiendo las líneas descritas a continuación, se obtiene tal estimador

$$\widehat{Var}(\widehat{S}(t)) = \widehat{S}(t)^2 \sum_{j:t_j \leqslant t} \frac{d_j}{r_j(r_j - d_j)}, \tag{2.25}$$

que a menudo se conoce como fórmula de Greenwood. El estimador del error estándar para $\widehat{S}(t)$ está dado por la raíz cuadrada de (2.25) [25].

Se denota con $\underline{\widehat{h}} = (\widehat{h}(t_1), \ldots, \widehat{h}(t_g))'$ (donde $\widehat{h}(t_j) = \dfrac{d_j}{r_j}$, $j = 1, \ldots, g$) al estimador de máxima verosimilitud de $\underline{h} = (h(t_1), \ldots, h(t_g))'$. Se tiene que $\underline{\widehat{h}}$ es asintóticamente $N(\underline{h}, \mathcal{I}^{-1}(\underline{h}))$, o equivalentemente $\sqrt{n}(\underline{\widehat{h}} - \underline{h})$ es asintóticamente $N(\underline{0}, n\mathcal{I}^{-1}(\underline{h}))$ en donde

$$\mathcal{I}(\underline{h}) = E\left(-\frac{\partial^2 l(\underline{h}; \underline{t})}{\partial \underline{h} \partial \underline{h}'}\right),$$

es la matriz de información de Fisher o matriz de información esperada que se puede estimar con la matriz de información observada

$$\begin{aligned} -\frac{\partial^2 l(\underline{h}; \underline{t})}{\partial \underline{h} \partial \underline{h}'}\bigg|_{\underline{h}=\underline{\widehat{h}}} &= \left(\left\{-\frac{\partial^2 l(\underline{h}; \underline{t})}{\partial h(t_i) \partial h(t_j)}\right\}_{ij}\right)\bigg|_{\underline{h}=\underline{\widehat{h}}} \\ &= \begin{cases} \dfrac{r_j}{\widehat{h}(t_j)(1 - \widehat{h}(t_j))} & \text{si} \quad i = j \\ 0 & \text{si} \quad i \neq j, \end{cases} \end{aligned}$$

por lo que asintóticamente $\widehat{h}(t_1), \ldots, \widehat{h}(t_g)$ son independientes y

$$Var(\widehat{h}(t_j)) = \frac{\widehat{h}(t_j)(1 - \widehat{h}(t_j))}{r_j} = \frac{d_j(r_j - d_j)}{r_j^3}, j = 1, \ldots, g. \tag{2.26}$$

Además, como $\log \widehat{S}(t) = \sum_{j:t_j \leqslant t} \log\left(1 - \widehat{h}(t_j)\right)$, para t fijo se tiene que asintóticamente

$$Var(\log \widehat{S}(t)) = \sum_{j:t_j \leqslant t} Var[\log(1 - \widehat{h}(t_j))].$$

Luego usando (A.3) y (2.26) se tiene que

$$Var[\log(1-\widehat{h}(t_j))] = \left(\frac{1}{1-\widehat{h}(t_j)}\right)^2 Var(\widehat{h}(t_j)) = \frac{d_j}{r_j(r_j-d_j)},$$

por lo tanto

$$Var(\log \widehat{S}(t)) = \sum_{j:t_j \leqslant t} \frac{d_j}{r_j(r_j-d_j)}.$$

Aplicando otra vez (A.3) del Apéndice A se sigue que

$$\begin{aligned} Var(\widehat{S}(t)) &= Var(\exp(\log \widehat{S}(t))) \\ &= [\exp(\log S(t))]^2 Var(\log \widehat{S}(t))) \\ &= S(t)^2 \sum_{j:t_j \leqslant t} \frac{d_j}{r_j(r_j-d_j)}, \end{aligned}$$

por lo tanto se tiene (2.25).

Los límites de confianza para $S(t)$, en el caso de datos con censura por la derecha, se pueden obtener usando la aproximación normal basada en $\widehat{S}(t)$, $\log \widehat{S}(t)$ o en $\log(-\log(\widehat{S}(t)))$. En el caso de $\widehat{S}(t)$, los límites de confianza correspondientes al intervalo de confianza con coeficiente de confianza $1-\alpha$ para $S(t)$ son

$$\widehat{S}(t) \pm z_{\alpha/2}[\widehat{Var}(\widehat{S}(t))]^{1/2},$$

en donde $z_{\alpha/2}$ es el $\alpha/2$ cuantil de la distribución normal estándar y $\widehat{Var}(\widehat{S}(t))$ está dada por (2.25).

Usando (A.3) se tiene que

$$\begin{aligned} Var(\log(-\log(\widehat{S}(t))) &= \left(\frac{1}{-\log(S(t))}\right)^2 Var(-\log(\widehat{S}(t)) \\ &= \left(\frac{1}{\log(S(t))}\right)^2 Var(\log(\widehat{S}(t)) \\ &= \left(\frac{1}{S(t)\log(S(t))}\right)^2 Var(\widehat{S}(t)), \end{aligned}$$

por lo que

$$\widehat{Var}(\log(-\log(\widehat{S}(t))) = \left(\frac{1}{\widehat{S}(t)\log(\widehat{S}(t))}\right)^2 \widehat{Var}(\widehat{S}(t)) \qquad (2.27)$$

donde $\widehat{Var}(\widehat{S}(t))$ está dada por (2.25).

Ahora, si se considera $\log(-\log(\widehat{S}(t)))$, los límites de confianza correspondientes al intervalo de confianza con coeficiente de confianza $1-\alpha$ para $S(t)$ son

$$\exp\left\{-\exp\left\{\log(-\log(\widehat{S}(t))) \mp z_{\alpha/2}[\widehat{Var}(\log(-\log(\widehat{S}(t))))]^{1/2}\right\}\right\},$$

en donde $z_{\alpha/2}$ es el $\alpha/2$ cuantil de la distribución normal estándar y

$$\widehat{Var}(\log(-\log(\widehat{S}(t))))$$

está dada por (2.27).

2.5.2. Estimador de Nelson-Aalen

El estimador de la función de riesgo acumulado correspondiente a (2.24) está dado por la integral de Riemann-Stieltjes como

$$\widehat{H}(t) = \int_0^t d\widehat{H}(u) = \int_0^t \frac{dN(u)}{Y(u)},$$

se asume que $Y(u) > 0$ para $0 \leqslant u \leqslant t$. Éste es llamado algunas veces la función de riesgo acumulativo empírica, pero es comúnmente más conocido como el estimador de Nelson-Aalen. En la notación usada para el estimador de Kaplan-Meier se tiene que

$$\widehat{H}(t) = \sum_{j:t_j \leqslant t} \frac{d_j}{r_j},$$

en donde $t_1, \ldots, t_k$ representan los distintos tiempos en que las fallas son observadas. Las gráficas de $\widehat{H}(\cdot)$ proporcionan información útil, por ejemplo, $H(\cdot)$ es lineal, si $h(\cdot)$ es constante [25].

El desarrollo de máxima verosimilitud que conduce a (2.25) también proporciona un estimador de la varianza asintótica para $\widehat{H}(t)$. Dado que asintóticamente $\widehat{h}(t_1), \ldots, \widehat{h}(t_g)$ son independientes (donde $\widehat{h}(t_j) = \dfrac{d_j}{r_j}$, $j = 1, \ldots, g$), por (2.26) se tiene que

$$\begin{aligned} Var(\widehat{H}(t)) &= Var\left(\sum_{j:t_j \leqslant t} \widehat{h}(t_j)\right) \\ &= \sum_{j:t_j \leqslant t} Var\left(\widehat{h}(t_j)\right) \\ &= \sum_{j:t_j \leqslant t} \frac{d_j(r_j - d_j)}{r_j^3}. \end{aligned}$$

2.6. Criterio de información de Akaike (CIA) y criterio de información Bayesiano (CIB)

La selección de un modelo entre un conjunto finito de modelos se puede llevar acabo mediante el criterio de información de Akaike (CIA) y mediante el criterio de información Bayesiano (CIB). El CIA es una medida de la bondad de ajuste de un modelo estadístico estimado. Este criterio se basa en el concepto de entropía, proporciona una medida relativa de la información perdida cuando se utiliza un modelo dado para describir la realidad y se puede decir que describe el equilibrio entre el sesgo y la varianza en la construcción del modelo, es decir, la precisión y complejidad del modelo. El CIA no es una prueba del modelo en el sentido de prueba de hipótesis, más bien es una prueba entre modelos, una herramienta para la selección de modelos. Dado un conjunto de datos, los diferentes modelos que compiten se pueden clasificar de acuerdo a su CIA, el modelo con el menor CIA será el mejor. En general, el CIA es

$$CIA = 2k - 2\log(L(\widehat{\underline{\phi}})),$$

donde k es el número de parámetros del modelo estadístico y $L(\widehat{\underline{\phi}})$ es el valor máximo de la función de verosimilitud para el modelo estadístico [21].

El CIB o Criterio Schwarz es un criterio para la selección de un modelo entre una clase de modelos paramétricos con diferente número de parámetros. Elegir un modelo que optimice el CIB es una forma de regularización. Este criterio está muy relacionado con el CIA. En el CIB, la sanción por parámetros adicionales es más fuerte que en el CIA. La fórmula para el CIB es

$$CIB = k\log n - 2\log(L(\widehat{\underline{\phi}})).$$

La metodología del CIA y del CIB intenta encontrar el modelo que explique mejor los datos con el mínimo de sus valores [21].

Capítulo 3

Modelos de regresión de supervivencia

El desarrollo presentado en el Capítulo 2 sólo involucra a la variable tiempo. Sin embargo, frecuentemente se desea comparar dos o más conjuntos de datos, algunas veces es mejor hacer la estimación de la función de supervivencia para cada conjunto de datos por separado y luego hacer una comparación cualitativa, ya sea directamente o mediante un resumen estadístico. También se pueden hacer comparaciones más sensibles o más complejas mediante modelos completos en los que el efecto de las covariables o variables explicativas se representa por medio de parámetros desconocidos [9]. Por otra parte, en muchos estudios el principal objetivo es entender y aprovechar la relación entre el tiempo de falla y las covariables [25].

En el presente capítulo se muestran algunos de los posibles modelos que se pueden usar para representar el efecto de variables explicativas en el tiempo de falla. Para esto supóngase que para cada individuo está definido un vector $\underline{x}$ de variables explicativas. Las componentes de $\underline{x}$ pueden representar varias características, que se piensa afectan el tiempo de falla, tales como tratamientos, indicadores de grupo, características individuales o condiciones ambientales.

Las variables explicativas se pueden clasificar como constantes (fijas), $\underline{x}$, o dependientes del tiempo, $\underline{x}(t)$. Un proceso de las covariables $X = \{\underline{x}(t)|t \geqslant 0\}$, que se desarrolla independientemente del proceso de tiempo de falla, se denomina externo, mientras que un proceso de las covariables, que se desarrolla dependiente del proceso de tiempo de falla, se denomina interno y su tratamiento requiere cuidado. Notar que las covariables constantes son externas [25].

3.1. Modelos de regresión paramétricos

Los modelos de regresión para tiempos de falla se pueden formular de diferentes formas. Cualquiera de los modelos paramétricos presentados en la Sección 2.4 se puede convertir en un modelo de regresión paramétrico especificando una relación entre los parámetros del modelo y las covariables. Sin embargo, frecuentemente sólo ciertos parámetros en una distribución de tiempo de falla se asumen dependientes de las covariables [25]. En las Subsecciones 3.1.2 y 3.1.3 se presentan dos modelos que se formulan de esta forma.

Otra forma frecuente de formular modelos es definir al vector $\underline{x}$ de variables explicativas de modo que $\underline{x} = \underline{0}$ corresponde a algún conjunto de condiciones estándar significativas, por ejemplo un tratamiento control. Luego los modelos se pueden desarrollar convenientemente en dos partes:

1. un modelo para la distribución del tiempo de falla cuando $\underline{x} = \underline{0}$;

2. una representación del cambio inducido por un $\underline{x} \neq \underline{0}$, frecuentemente en términos de alguna forma paramétrica [9].

Dos modelos que se pueden desarrollar de esta forma son el modelo de vida acelerada y el modelo de riesgo proporcional, los cuales se analizan en las Subsecciones 3.1.4 y 3.1.5, respectivamente.

3.1.1. La función de verosimilitud de datos que contienen covariables

El análisis de los modelos de regresión paramétricos se concentra en los métodos basados en la función de verosimilitud. A continuación se presenta la función de verosimilitud para datos que contienen covariables y están sujetos a censura por la derecha o por intervalo.

Datos censurados por la derecha

Primero, supóngase que se tienen datos que contienen covariables constantes y están sujetos a censura por la derecha. La función de verosimilitud se calcula de manera análoga a la función de verosimilitud de la Subsección 2.1.3, en este caso se consideran a las probabilidades, la f.d.p., la función de supervivencia y la función de riesgo condicionadas al valor de las covariables. Por lo que, para una variedad de mecanismos de censura por la derecha la

función de verosimilitud de las observaciones $(t_i, \delta_i, \underline{x}_i)$, $i = 1, \ldots, n$, cumple que

$$L \propto \prod_{i=1}^{n} f(t_i|\underline{x}_i)^{\delta_i} S(t_i|\underline{x}_i)^{1-\delta_i},$$

en donde δ_i es el indicador de censura para t_i.

Ahora, supóngase que se tienen datos que contienen covariables que varían en el tiempo, $\underline{x}(\cdot)$. Un enfoque conveniente para modelar con covariables que varían en el tiempo es mediante la función de riesgo, para la cual se puede permitir que dependa de la historia previa de las covariables. Sea $X(t) = \{\underline{x}(s)|0 \leqslant s \leqslant t\}$, que denota la historia de las covariables hasta el tiempo t, con $X(\infty) = X$. Un supuesto natural es que la función de riesgo para T dado X dependa sólo de $X(t)$; esto se denota como $h(t|X(t))$. En el caso de distribuciones continuas, la relación entre la función de riesgo y la función de supervivencia es [25]

$$S(t|X) = Pr(T > t|X) = \exp\left[-\int_0^t h(u|X)du\right], \tag{3.1}$$

luego la f.d.p. está dada por

$$f(t|X) = h(t|X)S(t|X) = h(t|X)\exp\left[-\int_0^t h(u|X)du\right]. \tag{3.2}$$

Considerando lo anterior, el desarrollo presentado en la Subsección 2.1.3 se puede extender. En este caso se le agrega $X_1(t), \ldots, X_n(t)$ a $\mathcal{H}(t)$. Por lo que, para una variedad de mecanismos de censura por la derecha, la función de verosimilitud de las observaciones $(t_i, \delta_i, X_i(t_i))$, $i = 1, \ldots, n$, cumple que

$$L \propto \prod_{i=1}^{n} f(t_i|X_i)^{\delta_i} S(t_i|X_i)^{1-\delta_i},$$

en donde δ_i es el indicador de censura para t_i.

Datos censurados por intervalo

En el caso de datos que contienen covariables constantes y están sujetos a censura por intervalo. La función de verosimilitud se calcula de manera análoga a la función de verosimilitud de la Sección 2.2, en este caso se supone que la elección de t_{ij} es condicionalmente independiente de la información de fallas más allá de $t_{i,j-1}$, dado $\mathcal{H}(t_{i,j-1})$ y $\underline{x}$, también se consideran a las probabilidades y a la f.d.a. condicionadas al valor de las covariables. Por lo

que para una variedad de mecanismos de censura por intervalo la función de verosimilitud de las observaciones $(t_{i,m_i-1}, t_{im_i}, \underline{x}_i)$, $i = 1, \ldots, n$, cumple que

$$L \propto \prod_{i=1}^{n} [F(t_{im_i}|\underline{x}_i) - F(t_{i,m_i-1}|\underline{x}_i)],$$

en donde $F(t|\underline{x}_i) = 1 - S(t|\underline{x}_i)$.

Ahora, supóngase que se tienen datos que contienen covariables que varían en el tiempo, $\underline{x}(\cdot)$. Considerando (3.1) y (3.2) se puede extender el desarrollo presentado en la Sección 2.2. En este caso $\mathcal{H}(t_{i,j-1})$ también contiene la información de $X_1(t_{i,j-1}), \ldots, X_n(t_{i,j-1})$. Por lo que para una variedad de mecanismos de censura por intervalo la función de verosimilitud de las observaciones $(t_{i,m_i-1}, t_{im_i}, X_i(t_{im_i}))$, $i = 1, \ldots, n$, cumple que

$$L \propto \prod_{i=1}^{n} [F(t_{im_i}|X_i) - F(t_{i,m_i-1}|X_i)],$$

en donde $F(t|X) = 1 - S(t|X)$.

3.1.2. Modelo de regresión exponencial

Suponga que en una población cada individuo tiene un tiempo de falla T y un vector $\underline{x} = (x_1, \ldots, x_p)'$ de covariables constantes. Entonces, por ejemplo, se puede asumir que dado $\underline{x}$ se tiene un modelo con distribución exponencial, i.e. la distribución de T es exponencial con función de supervivencia

$$S(t|\underline{x}) = \exp\{-\lambda(\underline{x})t\}, \text{ para } t \geqslant 0.$$

La especificación del modelo de regresión exponencial también involucra una forma funcional para $\lambda(\cdot)$. Una forma común es $\lambda(\underline{x}) = \exp\{\underline{\beta}'\underline{x}\}$, donde $\underline{\beta}$ $(p \times 1)$ es un vector de coeficientes de regresión, esta función tiene la propiedad de que $\lambda(\underline{x}) > 0$ para todos los vectores $\underline{\beta}$ y $\underline{x}$, lo cual es conveniente ya que el parámetro de la distribución exponencial es positivo [25].

Si se supone el modelo de regresión exponencial, para una variedad de mecanismos de censura por la derecha la función de verosimilitud de las observaciones $(t_i, \delta_i, \underline{x}_i)$, $i = 1, \ldots, n$, cumple que

$$L(\underline{\beta}) = \prod_{i=1}^{n} (\lambda(\underline{x}_i; \underline{\beta}) e^{-\lambda(\underline{x}_i;\underline{\beta})t})^{\delta_i} (e^{-\lambda(\underline{x}_i;\underline{\beta})t})^{1-\delta_i} = \prod_{i \in \mathscr{U}} \lambda(\underline{x}_i; \underline{\beta}) \prod_{i=1}^{n} e^{-\lambda(\underline{x}_i;\underline{\beta})t}$$

en donde $\lambda(\underline{x}_i;\underline{\beta})$ es el parámetro de la distribución del tiempo de falla del i-ésimo individuo y $\mathscr{U}$ denota al conjunto de los individuos que fallan. Luego la función $\log$ de verosimilitud está dada por

$$l(\underline{\beta}) = \sum_{i\in\mathscr{U}} \log(\lambda(\underline{x}_i;\underline{\beta})) - \sum_{i=1}^{n} \lambda(\underline{x}_i;\underline{\beta})t.$$

3.1.3. Modelo de regresión Weibull

Ahora, considérese la distribución Weibull con parámetro de escala λ y parámetro de forma θ. Se pueden formular modelos de regresión en los que se considere que λ y/o θ dependan de $\underline{x}$. Dado que λ y θ son positivos, nuevamente un par de especificaciones convenientes son $\lambda(\underline{x}) = \exp\{\underline{\beta}'\underline{x}\}$ y $\theta(\underline{x}) = \exp\{\underline{\gamma}'\underline{x}\}$ donde $\underline{\beta}$ y $\underline{\gamma}$ son vectores de coeficientes de regresión. Un modelo Weibull que es útil en muchas situaciones cumple que solamente λ depende de $\underline{x}$ y la función de supervivencia de T es [25]

$$S(t|\underline{x}) = \exp\{-(t/\lambda(\underline{x}))^{\theta}\}, \text{ para } t \geqslant 0.$$

3.1.4. Modelo de vida acelerada

Un modelo que ha sido ampliamente usado en el análisis de datos de supervivencia es el modelo de vida acelerada. Este modelo se define para variables explicativas constantes (modelo de vida acelerada simple) y para variables explicativas dependientes del tiempo como sigue.

Variables explicativas constantes

Supóngase que hay dos tratamientos representados por los valores 0 y 1 de una variable explicativa x. Sean $S_0(\cdot)$, $f_0(\cdot)$ y $h_0(\cdot)$ la función de supervivencia, la f.d.p. y la función de riesgo, respectivamente, en $x = 0$; en el modelo de vida acelerada simple o estándar hay una constante $\psi > 0$ tal que la función de supervivencia en $x = 1$ es

$$S_1(t) = S_0(\psi t),$$

por lo que la f.d.p. y la función de riesgo en $x = 1$ son

$$f_1(t) = -\frac{d}{dt}S_1(t) = -\psi S_0'(\psi t) = \psi f_0(\psi t)$$

y

$$h_1(t) = \frac{f_1(t)}{S_1(t)} = \frac{\psi f_0(\psi t)}{S_0(\psi t)} = \psi h_0(\psi t).$$

Una versión más fuerte es que cualquier individuo teniendo tiempo de supervivencia t bajo $x = 0$ tendrá tiempo de supervivencia t/ψ bajo $x = 1$, i.e. las correspondientes variables aleatorias están relacionas por $T_1 = T_0/\psi$ [9].

Más generalmente, con un vector de variables explicativas constantes $\underline{x}$, supóngase que hay una función positiva $\psi(\cdot)$ tal que la función de supervivencia, la f.d.p. y la función de riesgo son respectivamente

$$\begin{array}{rcl} S(t|\underline{x}) & = & S_0(t\psi(\underline{x})), \\ f(t|\underline{x}) & = & f_0(t\psi(\underline{x}))\psi(\underline{x}), \\ h(t|\underline{x}) & = & h_0(t\psi(\underline{x}))\psi(\underline{x}), \end{array} \tag{3.3}$$

en donde $S_0(\cdot)$ es la función de supervivencia en $\underline{x} = \underline{0}$ y $\psi(\underline{0}) = 1$.

Una representación en términos de variables aleatorias es

$$T = T_0/\psi(\underline{x}), \tag{3.4}$$

en donde T_0 tiene función de supervivencia $S_0(\cdot)$. Si $\mu_0 = E(\log T_0)$, se puede escribir esto como

$$\log T = \mu_0 - \log \psi(\underline{x}) + \varepsilon, \tag{3.5}$$

en donde ε es una variable aleatoria con media cero y con distribución que no depende de $\underline{x}$ [9].

De (3.3) se sigue que si $\underline{x}_1$ y $\underline{x}_2$ cumplen que $\psi(\underline{x}_1) < \psi(\underline{x}_2)$ entonces $S(t|\underline{x}_2) \leqslant S(t|\underline{x}_1)$ para $t \geqslant 0$.

Hasta ahora la función de supervivencia en $\underline{x} = \underline{0}$, $S_0(\cdot)$, no se ha especificado. Si se considera a $S_0(\cdot)$ igual a la función de supervivencia de alguna familia paramétrica de la Sección 2.4, se obtiene una familia especial de modelos de vida acelerada simple. Si, además, $\psi(\cdot)$ se especifica paramétricamente, se obtiene un modelo completamente paramétrico. La forma paramétrica de $\psi(\cdot)$ se denotará por $\psi(\cdot; \underline{\beta})$.

Dado que $\psi(\underline{x}; \underline{\beta}) > 0$ y $\psi(\underline{0}; \underline{\beta}) = 1$, un candidato natural es

$$\psi(\underline{x}; \underline{\beta}) = e^{\underline{\beta}'\underline{x}},$$

en donde $\underline{\beta}$ es un vector de parámetros. Entonces con esta elección de $\psi(\cdot; \underline{\beta})$ (3.5) puede ser escrita como

$$\log T = \mu_0 - \underline{\beta}'\underline{x} + \varepsilon,$$

que es un modelo de regresión lineal [9].

Variables explicativas dependientes del tiempo

Supóngase ahora que el vector de variables explicativas depende del tiempo, $\underline{x}(t)$. La esencia del modelo de vida acelerada es que el tiempo es contraído o ampliado relativamente a éste en $\underline{x} = \underline{0}$. Esto sugiere que para un individuo caracterizado por $\underline{x}(t)$, el tiempo $t^{(\underline{x})}$, evoluciona con respecto al tiempo $t^{(\underline{0})}$ para tal individuo estando en $\underline{x} = \underline{0}$ de acuerdo con

$$\frac{dt^{(\underline{x})}}{dt^{(\underline{0})}} = \frac{1}{\psi[\underline{x}(t^{(\underline{x})})]},$$

i.e.

$$t^{(\underline{0})} = \int_0^{t^{(\underline{x})}} \psi[\underline{x}(u)]du = \Psi(t^{(\underline{x})}),$$

así que el tiempo de falla está relacionado, en lugar de (3.4), por $T = \Psi^{-1}(T_0)$ [9].

Por lo tanto, en el modelo de vida acelerada la función de supervivencia, la f.d.p. y la función de riesgo son

$$\begin{array}{rcl} S(t|X) & = & S_0[\Psi(t)], \\ f(t|X) & = & \psi[\underline{x}(t)]f_0[\Psi(t)], \\ h(t|X) & = & \psi[\underline{x}(t)]h_0[\Psi(t)], \end{array} \tag{3.6}$$

en donde $X = \{\underline{x}(t)|t \geqslant 0\}$ y $\psi(\cdot)$ es una función positiva. Si $\underline{x}(\cdot)$ es constante sobre el tiempo, i.e. $\underline{x}(t) = \underline{x}$ para todo $t \geqslant 0$, entonces (3.6) llega a ser (3.3).

Nuevamente, se puede considerar a $S_0(\cdot)$ igual a la función de supervivencia de alguna familia paramétrica de la Sección 2.4 y una forma paramétrica para $\psi(\cdot)$, que se denotará por $\psi(\cdot;\underline{\beta})$, para tener un modelo paramétrico. Un procedimiento común es especificar a $\psi[\underline{x}(u)] = e^{\underline{\beta}'\underline{x}(u)}$ para todo $u \geqslant 0$, en tal caso

$$S(t|X) = S_0[\Psi(t)] = S_0\left[\int_0^t \psi[\underline{x}(u)]du\right] = S_0\left[\int_0^t e^{\underline{\beta}'\underline{x}(u)}du\right].$$

3.1.5. Modelo de riesgo proporcional

Una segunda familia de modelos que ha sido ampliamente usada en el análisis de datos de supervivencia se especifica mejor mediante la función de riesgo, éste es el modelo de riesgo proporcional el cual se define para variables explicativas constantes (modelo de riesgo proporcional simple) y para variables explicativas dependientes del tiempo como sigue.

Variables explicativas constantes

En el modelo de riesgo proporcional simple se supone que para un vector $\underline{x}$ de variables explicativas constantes la función de riesgo es

$$h(t|\underline{x}) = \psi(\underline{x})h_0(t), \text{ para } t \geqslant 0, \tag{3.7}$$

en donde $\psi(\cdot)$ es positiva, $\psi(\underline{0}) = 1$ y $h_0(\cdot)$ es la función de riesgo para un individuo bajo la condición estándar ($\underline{x} = \underline{0}$) que se denomina función de riesgo basal. Luego la función de supervivencia y la f.d.p. cumplen que

$$\begin{aligned} S(t|\underline{x}) &= \exp[-H(t|\underline{x})] \\ &= \exp\left[-\int_0^t h(u|\underline{x})du\right] \\ &= \exp\left[-\psi(\underline{x})\int_0^t h_0(u)du\right] \\ &= [\exp[-H_0(t)]]^{\psi(\underline{x})}, \end{aligned}$$

i.e.

$$S(t|\underline{x}) = [S_0(t)]^{\psi(\underline{x})} \tag{3.8}$$

y

$$f(t|\underline{x}) = -\frac{d}{dt}S(t|\underline{x}) = \psi(\underline{x})[S_0(t)]^{\psi(\underline{x})-1}f_0(t) = \psi(\underline{x})h_0(t)[S_0(t)]^{\psi(\underline{x})}.$$

Por lo que las funciones de supervivencia forman la familia Lehmann generada por $S_0(\cdot)$. A (3.7) se le llama el modelo de riesgo proporcional simple.

De (3.7) y (3.8) se sigue que si $\underline{x}_1$ y $\underline{x}_2$ cumplen que $\psi(\underline{x}_1) < \psi(\underline{x}_2)$ entonces $h(t|\underline{x}_1) \leqslant h(t|\underline{x}_2)$ para $t \geqslant 0$ y $S(t|\underline{x}_2) \leqslant S(t|\underline{x}_1)$ para $t \geqslant 0$.

Los modelos completamente paramétricos se obtienen considerando a $h_0(\cdot)$ igual a la función de riesgo de alguna de las familias de distribuciones de la Sección 2.4 y una forma paramétrica para $\psi(\cdot)$.

Tres parametrizaciones para $\psi(\cdot)$ son: la forma $\log$ lineal $\psi(\underline{x};\underline{\beta}) = e^{\underline{\beta}'\underline{x}}$, que por buenas razones ha llegado a ser la más popular, la forma lineal $\psi(\underline{x};\underline{\beta}) = 1 + \underline{\beta}'\underline{x}$, y el logístico, $\psi(\underline{x};\underline{\beta}) = \log(1 + e^{\underline{\beta}'\underline{x}})$. La discriminación entre estas formas puede lograrse ajustando una familia aumentada. Por ejemplo, la familia

$$\psi(\underline{x};\underline{\beta},\kappa) = (1 + \kappa\underline{\beta}'\underline{x})^{1/\kappa}$$

que incluye a los modelos lineal y $\log$ lineal, cuando $\kappa = 1$ y $\kappa \to 0$ respectivamente [9].

Relación con el modelo de vida acelerada simple

En el caso de variables explicativas constantes un aspecto de interés es saber cuando el modelo de riesgo proporcional simple con función de riesgo $h(t|\underline{x}) = \psi_{RP}(\underline{x})h_0(t)$ es también un modelo de vida acelerada simple. Para esto se necesita que exista una función $\psi_{VA}(\cdot)$ tal que

$$[S_0(t)]^{\psi_{RP}(\underline{x})} = S_0[t\psi_{VA}(\underline{x})], \forall t \geqslant 0 \text{ y } \forall \underline{x}.$$

Equivalentemente se tiene que

$$[S_0(\exp(\tau))]^{\psi_{RP}(\underline{x})} = S_0[\exp(\tau)\psi_{VA}(\underline{x})] \tag{3.9}$$

en donde $\tau = \log(t)$.

Aplicando $\log(-\log(\cdot))$ en ambos lados de (3.9) se tiene que

$$\log(-\log([S_0(\exp(\tau))]^{\psi_{RP}(\underline{x})})) = \log(-\log(S_0[\exp(\tau)\psi_{VA}(\underline{x})]))$$

i.e.

$$\begin{array}{l} \log(\psi_{RP}(\underline{x})) + \log(-\log(S_0(\exp(\tau)))) \\ = \log(-\log(S_0(\exp(\tau + \log(\psi_{VA}(\underline{x})))))). \end{array} \tag{3.10}$$

Considerando a

$$g_0(\tau) = log[-log(S_0(\exp(\tau)))] \qquad \text{y} \qquad \lambda(\underline{x}) = log(\psi_{VA}(\underline{x})), \tag{3.11}$$

se tiene que (3.10) es equivalente a

$$log(\psi_{RP}(\underline{x})) + g_0(\tau) = g_0[\tau + \lambda(\underline{x})].$$

Para que esto se cumpla para todo τ y para alguna $\lambda(\cdot)$ tal que $\lambda(\underline{x}) \neq 0 \ \forall \underline{x}$, i.e. $\psi_{VA}(\underline{x}) \neq 1$, se necesita que

$$g_0(\tau) = \kappa\tau + \alpha \qquad \text{y} \qquad \lambda(\underline{x}) = \kappa^{-1} log(\psi_{RP}(\underline{x})),$$

en donde α y κ son constantes.

Luego de (3.11) se tiene que

$$\begin{array}{rcl} S_0(t) & = & \exp(-\exp(g_0(\tau))) \\ & = & \exp(-\exp(\kappa\log(t) + \alpha)) \\ & = & \exp(-(t^\kappa \exp(\alpha))) \\ & = & \exp(-(t\rho)^\kappa), \end{array}$$

en donde $\rho = \exp(\alpha/\kappa)$ y

$$\psi_{RP}(\underline{x}) = [\psi_{VA}(\underline{x})]^\kappa.$$

Esto es, la distribución Weibull es la única distribución inicial para la cual el modelo de vida acelerada simple y el modelo de riesgo proporcional simple coinciden.

Por lo tanto, si $\psi_{RP}(\underline{x}) = [\psi_{VA}(\underline{x})]^{\kappa}$ los siguientes modelos coinciden:

1. Modelo de vida acelerada simple

$$\begin{aligned} f(t|\underline{x}) &= f_0(t\psi_{VA}(\underline{x}))\psi_{VA}(\underline{x}) \\ &= \kappa(\rho\psi_{VA}(\underline{x}))^{\kappa}t^{\kappa-1}\exp(-(t\psi_{VA}(\underline{x})\rho)^{\kappa}), \\ S(t|\underline{x}) &= S_0(t\psi_{VA}(\underline{x})) \\ &= \exp(-(t\psi_{VA}(\underline{x})\rho)^{\kappa}), \\ h(t|\underline{x}) &= h_0(t\psi_{VA}(\underline{x}))\psi_{VA}(\underline{x}) \\ &= \kappa(\rho\psi_{VA}(\underline{x}))^{\kappa}t^{\kappa-1}. \end{aligned}$$

2. Modelo de riesgo proporcional

$$\begin{aligned} f(t|\underline{x}) &= \psi_{RP}(\underline{x})[S_0(t)]^{\psi_{RP}(\underline{x})-1}f_0(t) \\ &= \psi_{RP}(\underline{x})\rho\kappa(\rho t)^{\kappa-1}\exp[(-\rho t)^{\kappa}\psi_{RP}(\underline{x})], \\ S(t|\underline{x}) &= [S_0(t)]^{\psi_{RP}(\underline{x})} \\ &= \exp[(-\rho t)^{\kappa}\psi_{RP}(\underline{x})], \\ h(t|\underline{x}) &= \psi_{RP}(\underline{x})h_0(t) \\ &= \psi_{RP}(\underline{x})\rho\kappa(\rho t)^{\kappa-1}. \end{aligned}$$

En particular si $\psi_{VA}(\underline{x};\underline{\beta}_{VA}) = \exp(\underline{\beta}'_{VA}\underline{x})$, entonces

$$\psi_{RP}(\underline{x};\underline{\beta}_{RP}) = \exp(\underline{\beta}'_{RP}\underline{x})$$

con $\underline{\beta}_{RP} = \kappa\underline{\beta}_{VA}$.

Variables explicativas dependientes del tiempo

La especificación (3.7) del modelo de riesgo proporcional simple se extiende inmediatamente cuando las variables explicativas dependen del tiempo como sigue

$$h(t|\underline{x}(t)) = \psi(\underline{x}(t))h_0(t) \text{ para } t \geqslant 0, \tag{3.12}$$

por lo que la función de supervivencia y la f.d.p. son

$$\begin{aligned} S(t|\underline{x}(t)) &= \exp[-H(t|\underline{x}(t))] \\ &= \exp\left[-\int_0^t h(u|\underline{x}(u))du\right] \\ &= \exp\left[-\int_0^t \psi(\underline{x}(u))h_0(u)du\right] \end{aligned}$$

y

$$\begin{aligned} f(t|\underline{x}(t)) &= h(t|\underline{x}(t))S(t|\underline{x}(t)) \\ &= \psi(\underline{x}(t))h_0(t)\exp\left[-\int_0^t \psi(\underline{x}(u))h_0(u)du\right]. \end{aligned}$$

Si $\underline{x}(\cdot)$ es constante sobre el tiempo, i.e. $\underline{x}(t) = \underline{x}$ para todo $t \geqslant 0$, entonces (3.12) llega a ser (3.7), i.e. se tiene el modelo de riesgo proporcional simple.

Se puede considerar a $h_0(\cdot)$ igual a la función de riesgo de alguna familia paramétrica de la Sección 2.4 y a una forma paramétrica para $\psi(\cdot)$, que se denotará por $\psi(\cdot;\underline{\beta})$, para tener un modelo paramétrico.

3.1.6. Modelo de riesgo aditivo

En el modelo de vida acelerada y en el modelo de riesgo proporcional las distribuciones de T para diferentes valores en las covariables, $\underline{x}_1$ y $\underline{x}_2$, que cumplen que $\psi(\underline{x}_1) < \psi(\underline{x}_2)$ son ordenadas en el sentido de que $S(t|\underline{x}_1) \leqslant S(t|\underline{x}_2)$ o $S(t|\underline{x}_1) \geqslant S(t|\underline{x}_2)$ para $t \geqslant 0$. Existen otros modelos que también cumplen esta propiedad, uno es el modelo de riesgo aditivo en el cual

$$h(t|\underline{x}) = \psi(\underline{x}) + h_0(t) \text{ para } t \geqslant 0,$$

en donde $\psi(\cdot)$ está restringida para que $\psi(\underline{0}) = 0$ y $\psi(\underline{x}) + h_0(t) \geqslant 0$ para todo $\underline{x}$ y $t \geqslant 0$, además $h_0(\cdot)$ es la función de riesgo para un individuo bajo la condición estándar ($\underline{x} = \underline{0}$). La función de supervivencia y la f.d.p. cumplen que

$$\begin{aligned} S(t|\underline{x}) &= \exp[-H(t|\underline{x})] \\ &= \exp\left[-\int_0^t h(u|\underline{x})du\right] \\ &= \exp\left[-\int_0^t (\psi(\underline{x}) + h_0(u))du\right] \\ &= \exp\left[-\psi(\underline{x})t - H_0(t)\right] \\ &= \exp\left[-\psi(\underline{x})t\right]\exp\left[-H_0(t)\right], \end{aligned}$$

i.e.

$$S(t|\underline{x}) = \exp\left[-\psi(\underline{x})t\right]S_0(t) \tag{3.13}$$

y

$$f(t|\underline{x}) = h(t|\underline{x})S(t|\underline{x}) = (\psi(\underline{x}) + h_0(t))\exp\left[-\psi(\underline{x})t\right]S_0(t).$$

De (3.13) se sigue que si $\underline{x}_1$ y $\underline{x}_2$ cumplen que $\psi(\underline{x}_1) < \psi(\underline{x}_2)$ entonces $S(t|\underline{x}_2) \leqslant S(t|\underline{x}_1)$ para $t \geqslant 0$.

Los modelos completamente paramétricos se obtienen considerando a $h_0(\cdot)$ igual a la función de riesgo de alguna de las familias de distribuciones de la Sección 2.4 y una forma paramétrica para $\psi(\cdot)$.

3.2. Modelo de riesgo proporcional semiparamétrico

Los modelos semiparamétricos también son ampliamente usados, estos especifican la dependencia de T en $\underline{x}$ paramétricamente, pero consideran arbitraria a la distribución real. El modelo de regresión de tiempo de falla semiparamétrico más conocido es el modelo de riesgo proporcional semiparamétrico, que toma a la función de riesgo de T dado $\underline{x}$ de la forma

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

donde $h_0(\cdot)$ es una función de riesgo basal arbitraria [25].

De manera más general, considérese al modelo de riesgo proporcional simple, es decir,

$$h(t|\underline{x}) = \psi(\underline{x};\underline{\beta})h_0(t), \text{ para } t \geqslant 0,$$

en donde el vector de variables explicativas, $\underline{x}$, es constante para cualquier individuo. Se asume por el momento que los tiempos de supervivencia tienen distribución continua y son registrados exactamente, por lo que no hay posibilidad de vínculos.

La función de verosimilitud del modelo de riesgo proporcional semiparamétrico

Ahora se considera inferencia sobre $\underline{\beta}$ cuando la función de riesgo $h_0(\cdot)$ es completamente desconocida. Primero se trata el caso en el que no hay censura. Sean $\tau_1 < \tau_2 < \ldots < \tau_n$ los tiempos de falla ordenados de los n individuos y sea $\mathscr{I}_j$ la etiqueta del individuo que falla en τ_j. Así $\mathscr{I}_j = i$ si y sólo si $t_i = \tau_j$. Sea $\mathscr{R}(\tau_j) = \{i|t_i \geqslant \tau_j\}$ el conjunto de individuos en riesgo justo antes de la j-ésima falla ordenada y r_j su tamaño [9]. Estas definiciones se ilustran en la Figura 3.1.

El principio básico para la derivación de la verosimilitud es: los $\{\tau_j\}$ y $\{\mathscr{I}_j\}$ son conjuntamente equivalentes a los datos originales, es decir, a los tiempos de falla desordenados, t_i. Cuando no se conoce $h_0(\cdot)$, los τ_j pueden proporcionar poca o ninguna información sobre $\underline{\beta}$, debido a que su distribución dependerá en gran medida de $h_0(\cdot)$. Por lo tanto, la atención se puede enfocar en los $\mathscr{I}_j$. En el caso presente (sin censura), la distribución conjunta $p(i_1, i_2, \ldots, i_n)$ sobre el conjunto de todas las posibles permutaciones de $(1, 2, \ldots, n)$ se puede derivar explícitamente [9].

La probabilidad condicional de que $\mathscr{I}_j = i$ dada toda la historia hasta el j-ésimo tiempo de falla ordenado τ_j, $\mathscr{H}_j = \{\tau_1, \tau_2, \ldots, \tau_{j-1}, \tau_j, i_1, i_2, \ldots, i_{j-1}\}$,

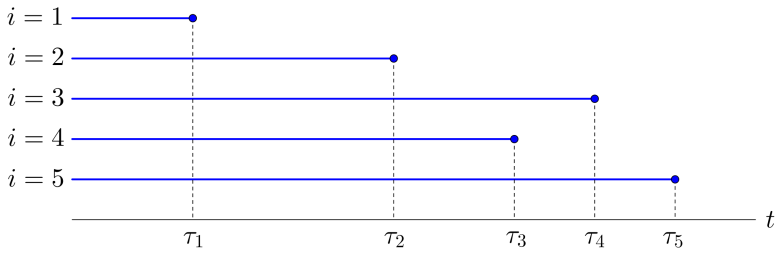

Figura 3.1: Se muestran las fallas(•) de cinco individuos (no hay censura). Los instantes de falla son $\tau_1, \tau_2, \tau_3, \tau_4, \tau_5$ y los conjuntos de riesgo son $\mathscr{R}(\tau_1) = \{1, 2, 3, 4, 5\}$, $\mathscr{R}(\tau_2) = \{2, 3, 4, 5\}$, $\mathscr{R}(\tau_3) = \{3, 4, 5\}$, $\mathscr{R}(\tau_4) = \{3, 5\}$ y $\mathscr{R}(\tau_5) = \{5\}$.

es la probabilidad condicional de que i falle en τ_j dado que un individuo del conjunto de riesgo $\mathscr{R}(\tau_j)$ falla en τ_j, es decir,

$$\begin{aligned} Pr(\mathscr{I}_j = i|\mathscr{H}_j) &= Pr(t_i = \tau_j| \text{ un individuo de } \mathscr{R}(\tau_j) \text{ falla en } \tau_j) \\ &= \frac{h(\tau_j|\underline{x}_i)}{\sum\limits_{k\in\mathscr{R}(\tau_j)} h(\tau_j|\underline{x}_k)}. \end{aligned}$$

Como se está suponiendo el modelo de riesgo proporcional simple se sigue que

$$\begin{aligned} Pr(\mathscr{I}_j = i|\mathscr{H}_j) &= \frac{h_0(\tau_j)\psi(\underline{x}_i;\underline{\beta})}{\sum\limits_{k\in\mathscr{R}(\tau_j)} h_0(\tau_j)\psi(\underline{x}_k;\underline{\beta})} \\ &= \frac{\psi(\underline{x}_i;\underline{\beta})}{\sum\limits_{k\in\mathscr{R}(\tau_j)} \psi(\underline{x}_k;\underline{\beta})} \\ &= \frac{\psi(i)}{\sum\limits_{k\in\mathscr{R}(\tau_j)} \psi(k)}, \end{aligned} \tag{3.14}$$

en donde $\psi(k)$ denota $\psi(\underline{x}_k, \underline{\beta})$, esto es, el factor $\psi(\cdot;\underline{\beta})$ para el k-ésimo sujeto.

A pesar de que (3.14) se deriva como la probabilidad condicional de $\mathscr{I}_j = i$ dada toda la historia $\mathscr{H}_j$, es condicionalmente independiente de los τ_1, τ_2, …, τ_{j-1}, τ_j. Por lo tanto es igual $Pr(\mathscr{I}_j = i|i_1, i_2, \ldots, i_{j-1})$, la distribución condicional de $\mathscr{I}_j$ dado solamente $\mathscr{I}_1 = i_1$, $\mathscr{I}_2 = i_2$, …, $\mathscr{I}_{j-1} = i_{j-1}$. Así,

la distribución conjunta de $Pr(i_1, i_2, \ldots, i_n)$ se puede obtener como sigue:

$$\begin{aligned} Pr(i_1, i_2, \ldots, i_n) &= \prod_{j=1}^{n} Pr(\mathscr{I}_j = i_j | i_1, i_2, \ldots, i_{j-1}) \\ &= \prod_{j=1}^{n} \frac{\psi(i_j)}{\sum_{k \in \mathscr{R}(\tau_j)} \psi(k)}. \end{aligned}$$

En el caso de presencia de censura se aplica un argumento similar si se puede asumir que las censuras sólo pueden ocurrir inmediatamente después de las fallas. Este requisito está ligeramente en desacuerdo con el modelo en el cual los tiempos censurados son contantes fijas, pero por lo general se puede ver como una aproximación razonable, ya que la información sobre $\underline{\beta}$ contribuida por un tiempo censurado observado exacto c_i generalmente será pequeña [9].

Supóngase que hay d fallas observadas en una muestra de tamaño n ($d \leqslant n$), sean $\tau_1 < \tau_2 < \ldots < \tau_d$ los tiempos de falla observados ordenados. Como en el caso sin censura, sea $\mathscr{I}_j = i$ si el sujeto i falla en τ_j, y sea $\mathscr{R}(\tau_j) = \{i | t_i \geqslant \tau_j\}$ el correspondiente conjunto de riesgo con tamaño r_j. La ecuación (3.14) sigue exactamente como antes, donde $\mathscr{H}_j$ ahora incluye las censuras en $(0, \tau_j)$ así como las fallas, y el hecho de que no pueda ocurrir censura en (τ_{j-1}, τ_j) que garantiza al conjunto de riesgo $\mathscr{R}(\tau_j)$, y así la expresión (3.14), no depende de τ_j [9].

La combinación de estas probabilidades condicionales dan la verosimilitud general

$$L = \prod_{j=1}^{d} \frac{\psi(i_j)}{\sum_{k \in \mathscr{R}(\tau_j)} \psi(k)}, \tag{3.15}$$

luego la función $\log$ de la verosimilitud es

$$l = \sum_{j=1}^{d} \left\{ \log(\psi(i_j)) - \log \left(\sum_{k \in \mathscr{R}(\tau_j)} \psi(k) \right) \right\}.$$

Con un ligero abuso de notación, (3.15) puede ser expresado en términos de los tiempos de falla observados no ordenados t_i, con correspondientes conjuntos de riesgo $\mathscr{R}(t_i) = \mathscr{R}_i$, como

$$L = \prod_{i \in \mathscr{U}} \frac{\psi(i)}{\sum_{k \in \mathscr{R}_i} \psi(k)},$$

en donde $\mathscr{U}$ denota al conjunto de los individuos que fallan. Luego la función $\log$ de la verosimilitud es

$$l = \sum_{i \in \mathscr{U}} l_i, \tag{3.16}$$

en donde $l_i = \log[\psi(i)] - \log\left(\sum_{k \in \mathscr{R}_i} \psi(k)\right)$.

Se denomina a (3.15) una verosimilitud más que una probabilidad, ya que se han omitido los términos que determinan qué individuos deben ser censurados entre los sobrevivientes de cada conjunto de riesgo. Mientras que el mecanismo de censura en sí mismo no dependa de $\underline{\beta}$, estos términos no dependerán funcionalmente de $\underline{\beta}$ y se puede ignorar el mecanismo de censura con el propósito de inferir con la verosimilitud sobre $\underline{\beta}$ [9].

Las derivadas de la función $\log$ de verosimilitud

La forma funcional de $\psi(\underline{x}; \underline{\beta})$ no es esencial para derivar (3.16), basta suponer que para todo i, $\psi(i)$ tiene primeras y segundas derivadas las cuales se denotarán por

$$\frac{\partial}{\partial \beta_r}\psi(i) = \psi_r(i), \qquad \frac{\partial^2}{\partial \beta_s \partial \beta_r}\psi(i) = \psi_{sr}(i).$$

Notar que en realidad no se necesita asumir que $\psi(\cdot)$ puede ser expresada explícitamente como una función de un vector covariable, aunque casi siempre lo será. Bastará con que el factor $\psi(i)$ de cada sujeto i sea expresado como una función de los parámetros $\underline{\beta}$ [9].

La primera derivada de la función $\log$ de verosimilitud (3.16) está dada por

$$\frac{\partial l}{\partial \beta_r} = \sum_{i \in \mathscr{U}} \frac{\partial l_i}{\partial \beta_r},$$

en donde

$$\frac{\partial l_i}{\partial \beta_r} = \frac{\psi_r(i)}{\psi(i)} - \frac{\sum_{k \in \mathscr{R}_i} \psi_r(k)}{\sum_{k \in \mathscr{R}_i} \psi(k)}.$$

Además, la segunda derivada de la función $\log$ de verosimilitud (3.16) está dada por

$$\frac{\partial^2 l}{\partial \beta_s \partial \beta_r} = \sum_{i \in \mathscr{U}} \frac{\partial^2 l_i}{\partial \beta_s \partial \beta_r},$$

en donde

$$\frac{\partial^2 l_i}{\partial\beta_s\partial\beta_r} = \frac{\psi_{sr}(i)}{\psi(i)} - \frac{\psi_r(i)\psi_s(i)}{[\psi(i)]^2} - \frac{\sum_{k\in\mathscr{R}_i}\psi_{sr}(k)}{\sum_{k\in\mathscr{R}_i}\psi(k)} + \frac{\sum_{k\in\mathscr{R}_i}\psi_r(k)\sum_{k\in\mathscr{R}_i}\psi_s(k)}{\left(\sum_{k\in\mathscr{R}_i}\psi(k)\right)^2}.$$

Parametrización $\log$ **lineal de** $\psi(\cdot)$

Si $\psi(\cdot)$ toma la forma $\log$ lineal, i.e. $\psi(\underline{x};\underline{\beta}) = e^{\underline{\beta}'\underline{x}}$, se tiene que $\psi_r(i) = x_{ir}\psi(i)$ y $\psi_{sr}(i) = x_{ir}x_{is}\psi(i)$, en donde x_{ir} denota el valor de la r-ésima componente del vector de variables explicativas, $\underline{x}$, en el i-ésimo individuo. En este caso se tiene que

$$\frac{\partial l_i}{\partial\beta_r} = x_{ir} - A_{ir}(\underline{\beta}), \tag{3.17}$$

en donde

$$A_{ir}(\underline{\beta}) = \frac{\sum_{k\in\mathscr{R}_i} x_{kr}e^{\underline{\beta}'\underline{x}_k}}{\sum_{k\in\mathscr{R}_i} e^{\underline{\beta}'\underline{x}_k}}.$$

Notar que (3.17) es la diferencia entre el valor de la variable explicativa en el sujeto que fallo y el promedio ponderado de la misma variable sobre el correspondiente conjunto de riesgo. También se tiene que

$$\begin{aligned}\frac{\partial^2 l_i}{\partial\beta_s\partial\beta_r} &= -\frac{\sum_{k\in\mathscr{R}_i} x_{kr}x_{ks}e^{\underline{\beta}'\underline{x}_k}}{\sum_{k\in\mathscr{R}_i} e^{\underline{\beta}'\underline{x}_k}} + \frac{\sum_{k\in\mathscr{R}_i} x_{kr}e^{\underline{\beta}'\underline{x}_k}\sum_{k\in\mathscr{R}_i} x_{ks}e^{\underline{\beta}'\underline{x}_k}}{\left(\sum_{k\in\mathscr{R}_i} e^{\underline{\beta}'\underline{x}_k}\right)^2} \\ &= -\frac{\sum_{k\in\mathscr{R}_i} x_{kr}x_{ks}e^{\underline{\beta}'\underline{x}_k}}{\sum_{k\in\mathscr{R}_i} e^{\underline{\beta}'\underline{x}_k}} + A_{ir}(\underline{\beta})A_{is}(\underline{\beta}).\end{aligned}$$

La inferencia basada en la función de verosimilitud (Sección 2.3) también se aplica aquí [9].

Capítulo 4

Caso de estudio: la deserción en las licenciaturas de la FCFM-BUAP

La deserción se entiende como la interrupción o desvinculación del proceso académico-institucional que lleva a cabo el estudiante. Esta ha sido un tema estudiado por diferentes autores e Instituciones de Educación Superior (IES). El tema ha tomado un lugar importante dado que no tiene sentido realizar un esfuerzo significativo por aumentar la cobertura, calidad y equidad en educación superior, sin controlar la deserción y su problemática multicausal y compleja. De esta manera, el emprendimiento de este tipo de estudios aporta a la comprensión del fenómeno y permite generar estrategias de retención estudiantil al interior de las IES y por parte del Estado [28].

La deserción se puede clasificar de diferentes maneras, por ejemplo, Vásquez, Castaño, Gallón y Gómez [43] identifican tres tipos de deserción:

1. Deserción precoz: el estudiante abandona el programa antes de comenzar habiendo sido aceptado.
2. Deserción temprana: el estudiante abandona el programa durante los primeros cuatro semestres.
3. Deserción tardía: entendida como abandono desde el quinto semestre en adelante.

Mientras que Stratton, O'Toole, y Wetzel [38] hacen una separación entre:

- deserción total (*dropout*): periodos de deserción a largo plazo o permanentes, y

- deserción parcial (*stopout*): periodos de deserción a corto plazo.

Es importante diferenciar la deserción total de la parcial ya que pueden tener diferentes factores causales y al no hacer esta distinción las investigaciones pueden identificar incorrectamente los factores asociados con deserción total. Para efectos prácticos, en algunos estudios se considera desertor a cualquier estudiante que abandone el programa o la institución por dos o más semestres consecutivos ([7], [28], [38], [42], [43]).

Una forma de clasificar los modelos es según el tipo de variables explicativas que juegan un papel en la explicación de la deserción. Las variables se separan en cuatro grupos específicos según la naturaleza de éstas ([7], [28], [43]):

- Individuales: Factores asociados específicamente al individuo.
- Socioeconómicos: Factores asociados a la situación económica, financiera y estrato social del estudiante.
- Académicos: Factores relacionados a la historia estudiantil de un individuo y su rendimiento en la universidad.
- Institucionales: Factores pertinentes a la Institución de Educación Superior específica a la que asiste un estudiante.

Tinto [41] plantea un modelo de deserción que denomina ser “longitudinal e interaccionista”. Es longitudinal en el sentido que sigue a un estudiante desde su ingreso a la institución de educación superior hasta que éste tome la decisión de desertar o termine satisfactoriamente su programa de estudios. Es interaccionista porque reconoce la importancia de la interacción de las variables explicativas del modelo, sin dar preferencia a un grupo particular, donde los factores de la institución académica son importantes en la explicación de la deserción, pero interactúan con otro tipo de variables para dar una explicación del fenómeno [28].

Algunos hallazgos de Tinto [40] muestran que cuanto más firme es el propósito personal de tener una carrera universitaria, mayor es la probabilidad de lograr la meta. Sin embargo, tales propósitos iníciales no son inalterables; cambian a lo largo de la experiencia. Por lo tanto, el mayor peso en la decisión de abandonar o proseguir recae sobre lo que ocurre una vez que el estudiante está adentro. Es decir, lo que ocurre “antes” del ingreso es importante, pero lo es más aquello que acontece “durante” la estadía del joven en la universidad [35].

Rodríguez y Leyva [33] reportan que en México, los pocos estudios sobre el comportamiento del fenómeno han encontrado que parecen existir causas que se pueden identificar como universales, las cuales son: las presiones económicas familiares, las dificultades de integración familiar, la reprobación escolar reincidente, problemas de salud, la edad de ingreso, y el traslape de horarios estudios-trabajo. En la Tabla 4.1 se muestran algunos de los factores causales que se han utilizado en diferentes estudios de deserción.

El estudio de Silva [35] establece que el primer año universitario constituye un tramo crítico que influye significativamente en una trayectoria exitosa o en una irregular y, por supuesto, en el abandono escolar. Además, durante el primer año ocurre la mayor incidencia de deserción o abandono escolar y existe un serio problema de rezago, debido frecuentemente a la reprobación. Estas evidencias revelan la importancia que tiene este período escolar por lo que es indispensable comprender la dinámica de este período escolar y diseñar estrategias para mejorarlo.

4.1. La deserción en las licenciaturas de la FCFM-BUAP

La Benemérita Universidad Autónoma de Puebla (BUAP), cuyas raíces se remontan al siglo XVI, constituye un gran pilar de la educación superior y la investigación científica en la región, y ocupa un destacado sitio entre las universidades públicas del país, gracias al esfuerzo conjunto de todos los miembros de la institución [16].

En la década de los 70, en la BUAP se impuso un modelo de Universidad Crítica, Democrática y Popular que fortaleció la investigación científica y la vinculación con los sectores más necesitados de la sociedad. Se creó el Instituto de Ciencias, se consolidó la Escuela de Físico Matemáticas y nacieron los primeros estudios de posgrado: maestría y doctorado en Física [16]. Actualmente, la oferta educativa de la Facultad de Ciencias Físico Matemáticas (FCFM) consta de:

- Licenciatura en Actuaría (LA)
- Licenciatura en Física (LF)
- Licenciatura en Física Aplicada (LFA)
- Licenciatura en Matemáticas (LM)
- Licenciatura en Matemáticas Aplicadas (LMA)

Individuales:	**Académicos:**	**Socioeconómico:**	**Institucionales:**
▪ Edad ▪ Género ▪ Estado civil ▪ Tener hijos ▪ Número de hermanos ▪ Posición entre hermanos ▪ Hermanos en educación superior ▪ Tipo de vivienda ▪ Ubicación de la vivienda ▪ Pertenece a una Etnia indígena ▪ Vivir con los padres ▪ Vivir en un hogar con padres divorciados	▪ Orientación ▪ Inicio inmediato ▪ Calificación del examen de admisión ▪ Primera opción ▪ Promedio ▪ Créditos cursados por semestre ▪ Calidad del esfuerzo del estudiante ▪ Actividades extracurriculares ▪ Servicios de apoyo ▪ Experiencia académica anterior ▪ Deserciones previas durante la trayectoria académica ▪ Educación preuniversitaria ▪ Haber estudiado en colegio público	▪ Estrato Socioeconómico ▪ Dependencia ▪ Personas a cargo ▪ Empleo estudiante ▪ Educación padres ▪ Ocupación padres	▪ Ritmo académico ▪ Procesos de admisión ▪ Relación con profesores ▪ Relación con compañeros ▪ Recursos ▪ Manejo pedagógico de los docentes

Tabla 4.1: Factores causales de deserción utilizados en diferentes estudios. Fuente: Elaboración propia con base en [27], [28], [35], [36], [37], [42], [43].

Lic.	2006	2007	2008	2009	2010	2011	2012	2013
LA					25 %	22 %	13 %	20 %
LM	39 %	27 %	36 %	28 %	57 %	32 %	33 %	43 %
LMA	34 %	35 %	42 %	26 %	30 %	37 %	44 %	43 %

Tabla 4.2: Porcentaje de deserción en el primer año de la LA, la LM y la LMA. Fuente: Elaboración propia a partir de información proporcionada por la FCFM-BUAP.

- Maestría en Ciencias Física Aplicada
- Maestría en Ciencias Matemáticas
- Maestría en Educación Matemática
- Doctorado en Ciencias Física Aplicada
- Doctorado en Ciencias Matemáticas.

Las licenciaturas de la FCFM-BUAP no están excentas del problema de la deserción escolar, hay alumnos desertores en los diferentes semestres impartidos. El periodo en el que se presenta la mayor deserción es el primer año escolar. Desde el 2006 hasta el 2013 el porcentaje de deserción en el primer año de la LM ha variado entre el 27 % y el 57 % y en la LMA entre el 26 % y el 44 %, mientras que en el caso de la LA la deserción varía entre el 13 % y el 25 % como se muestra en la Tabla 4.2. La primera generación de la LA ingresó en 2010 por lo que sólo se muestran las generaciones 2010 a la 2013.

4.2. Descripción del análisis de deserción para las licenciaturas de la FCFM-BUAP

En el análisis de deserción en las licenciaturas de la FCFM-BUAP se considera a la deserción como el abandono de la licenciatura de la FCFM. Por lo tanto, el seguimiento que se le hace al estudiante comienza desde que se matricula en una licenciatura de la FCFM hasta que sale de ésta. No se considera el hecho de que el estudiante continúe sus estudios en otra licenciatura de la FCFM, de la BUAP, de otra institución o si deja de estudiar por completo. Se supone que la deserción del estudiante es una decisión voluntaria y se reconoce que un estudiante ha desertado cuando permanece un semestre sin matricularse en la licenciatura que estaba cursando. El estudio sólo considera a la deserción total.

Covariable	Categorías
Género	Femenino (0) y Masculino (1)
Vive	Padre/Madre (0), Solo (1), Amigos (2), Esposo/Hijos (4) y Otros (3)
FinBach	Público (0) y Privado (1)
TipoBach	General (0) y Especializado (1)
MatRep	Ninguna (0), Menos de 3 (1), De 3 a 6 (2) y Más de 6 (3)
OpCarrera	Primera opción (0) y Segunda opción (1)
SosEst	Recursos familiares (0), Recursos propios (1), Ambos (2) y Otros (3)
Trabajo	No (0) y Si (1)
RecSem	Insuficientes (0), Suficientes (1) y Excelentes (2)

Tabla 4.3: Covariables con sus categorías y valores asignados.

El tiempo de observación considerado es del segundo semestre de 2009 al primer semestre de 2015, por lo que se consideran a las generaciones que ingresaron en 2009, 2010, 2011, 2012, 2013 y 2014.

Uno de los objetivos del análisis es determinar factores causales o indicadores de mayor deserción; para esto se ajusta un modelo interaccionista, es decir, se consideran covariables sin dar preferencia a ninguna en especial. Las covariables consideradas son: puntaje de ingreso (Puntaje), autoestima (Autoestima), hábitos de estudio (HabEstudio), razonamiento científico (Lawson), comprensión lectora (THLB), estilos de aprendizaje (Activo, Reflexivo, Teórico y Pragmático), género (Género), con quien vive (Vive), financiamiento del bachillerato de procedencia (FinBach), tipo de bachillerato de procedencia (TipoBach), materias reprobadas (MatRep), opción de carrera (OpCarrera), sostén de estudios (SosEst), trabajo (Trabajo) y recursos semanales (RecSem). Algunas covariables son categóricas por lo que se les asignó un valor a cada categoría como se muestra en la Tabla 4.3.

Las covariables consideradas son características del alumno obtenidas mediante las siguientes pruebas aplicadas a los estudiantes de nuevo ingreso de la FCFM (Tabla 4.4):

- Exámenes de admisión de la BUAP.
- Inventario de Autoestima de Coopersmith.
- Cuestionario de Hábitos de Estudio.
- Prueba de Lawson.

- Test de habilidades Lecto-Comprensivas Básicas.
- Cuestionario Honey-Alonso de Estilos de Aprendizaje.
- Cuestionario para alumnos de nuevo ingreso.

Estas pruebas se describen en la Sección 4.3.

Las pruebas aplicadas, excepto los exámenes de admisión de la BUAP, fueron elegidas en el proyecto Perfil Psicopedagógico de ingreso a las licenciaturas de la BUAP, desarrollado por: Dr. Osbaldo Germán Quiroz Romero, M.C. Ana Elena Posada, Dra. Olga Leticia Fuchs Gómez, Dra. Margarita Campos Méndez, M.C. Ana Elena Posada Sáchez y Dra. Eugenia Erica Vera Cervantes. La recopilación de la información utilizada en el análisis fue realizada por las coordinadoras de tutores de la FCFM-BUAP: Dra. Olga Leticia Fuchs Gómez y M.C. María Guadalupe Raggi Cárdenas.

Un limitante de este análisis es el hecho de que algunas covariables que varían en el tiempo son consideradas constantes.

Se aplica la metodología del Análisis de Supervivencia al problema de deserción escolar universitaria en las licenciaturas en Actuaría, Matemáticas y Matemáticas Aplicadas impartidas en la FCFM-BUAP. Se analiza la deserción de los alumnos de las licenciaturas mencionadas utilizando modelos no paramétricos, paramétricos y el modelo de riesgo proporcional semiparamétrico. El objetivo es obtener un modelo que permita pronosticar los tiempos de deserción en futuras generaciones e identificar las variables indicadoras de mayor riesgo que permitan tomar medidas de prevención adecuadas que disminuyan la deserción en las licenciaturas estudiadas.

En el caso bajo estudio la población objetivo consta de los alumnos que ingresaron a la FCFM entre los años 2009 y 2014. Para medir el tiempo de deserción para cada alumno se considera como tiempo origen la fecha de su ingreso a la FCFM, la escala para medir el paso del tiempo es el semestre y el significado de fracaso es la deserción.

El análisis estadístico se realiza con los paquetes *survival* y *fitdistrplus* del *software* R. El paquete *survival* permite calcular el estimador de Kaplan-Meier y estimar los parámetros del modelo de riesgo proporcional semiparamétrico. Mientras que el paquete *fitdistrplus* calcula estimadores de máxima verosimilitud.

El análisis de la deserción de los alumnos de la FCFM consta de dos etapas, en la primera no se incluyen covariables pero en la segunda sí.

Prueba:	**Factor causal medido:**	**Covariable:**
Exámenes de admisión de la BUAP	Puntaje de ingreso	Puntaje
Inventario de Autoestima de Coopersmith	Autoestima	Autoestima
Cuestionario de Hábitos de Estudio	Hábitos de estudio	HabEstudio
Prueba de Lawson	Razonamiento Científico	Lawson
Test de habilidades Lecto-Comprensivas Básicas	Comprensión lectora	THLB
Cuestionario Honey-Alonso de Estilos de Aprendizaje	Estilos de Aprendizaje	Activo
		Reflexivo
		Teórico
		Pragmático
Cuestionario para alumnos de nuevo ingreso	Género	Género
	Con quien vive	Vive
	Financiamiento del bachillerato de procedencia	FinBach
	Tipo de bachillerato de procedencia	TipoBach
	Materias reprobadas durante el bachillerato	MatRep
	Opción carrera	OpCarrera
	Sostén de estudios	SosEst
	Trabajo	Trabajo
	Recursos semanales	RecSem

Tabla 4.4: Pruebas y factores causales de deserción utilizados en el análisis de deserción de las licenciaturas de la FCFM-BUAP.

En la primera parte del estudio, se consideran censurados por la derecha a los estudiantes activos y los que tiene el 100 % de los créditos (titulados y no titulados), y censurados por intervalo a los estudiantes que desertaron, ya que sólo se sabe el semestre en el que desertaron. Debido a que el *software* R no estima a los parámetros del modelo de riesgo proporcional semiparamétrico con datos censurados por intervalo, en la segunda parte del análisis los datos censurados por intervalo se consideran tiempos realizados con tiempo de deserción igual al promedio de los extremos del intervalo de censura.

4.3. Pruebas aplicadas en la FCFM-BUAP

En esta sección se describen las pruebas aplicadas a los estudiantes de nuevo ingreso de la FCFM.

4.3.1. Exámenes de admisión de la BUAP

El proceso de admisión de la BUAP consta de 4 etapas [31]:

1. Registro en internet y selección de carrera.

2. Entrega del formato de asignación de examen.

3. Aplicación de la Prueba de Aptitud Académica y la Prueba por Área de Conocimiento.

   Todos los aspirantes de nivel Licenciatura y de Técnico Superior Universitario deben realizar dos exámenes como requisito de ingreso:

   - El primer examen consistente en la Prueba de Aptitud Académica (PAA). Dentro de ésta se realizará una sección llamada ESLAT, como instrumento para medir la competencia de inglés.
   - El segundo examen consiste en la Prueba por Área de Conocimiento (PAC).

4. Inscripción.

   Son candidatos a inscripción aquellos aspirantes que hayan obtenido un puntaje suficiente dentro de los cupos definidos por los Consejos de Unidad Académica correspondientes en la carrera elegida.

4.3.2. Inventario de Autoestima de Coopersmith

Coopersmith [8] define la autoestima como “el juicio personal de valía, que es expresado en las actitudes que el individuo toma hacia sí mismo. Es una experiencia subjetiva que se transmite a los demás por reportes verbales o conducta manifiesta”.

Debido a la gran influencia que la autoestima puede tener en la vida de las personas, se han desarrollado diferentes instrumentos con el fin de evaluarla. Coopersmith comienza en 1959 un estudio sobre la autoestima, y en 1967 publica una escala de medición de autoestima para niños, que es ampliamente utilizada. Coopersmith utilizó esta prueba como base para el desarrollo de la versión para adultos ([8], [24]).

El Inventario de Autoestima de Coopersmith (IAC) para niños consta de 58 declaraciones de respuesta dicotómica: igual que yo (si la declaración describe como se siente usualmente) o distinto a mi (si la declaración no describe como se siente usualmente). El inventario está referido a la percepción del estudiante en las siguientes áreas: autoestima general (26 declaraciones), autoestima social (8 declaraciones), autoestima familiar (en relación al hogar, 8 declaraciones), autoestima escolar-académica (8 declaraciones) y una escala de mentira (8 declaraciones) [6].

- Autoestima general: Corresponde al nivel de aceptación con el que la persona valora su conducta auto descriptiva.
- Autoestima social: Corresponde al nivel de aceptación con el que la persona valora su conducta en relación a sus pares.
- Autoestima familiar: Corresponde al nivel de aceptación con el que la persona valora su conducta en relación a su contexto familiar.
- Autoestima escolar-académica: Corresponde al nivel de aceptación con el que la persona valora su conducta en relación a su ámbito escolar.

El IAC tiene una Pauta de corrección en la cual se especifica qué declaraciones, con su respectiva respuesta, contribuyen al Puntaje Bruto (PB) de cada área medida. De cada declaración, sólo una respuesta contribuye 2 puntos al PB de una área específica, la otra respuesta no contribuye. Para obtener un indicador de la apreciación global que el sujeto tiene de si mismo se suman todos los PB, excepto el de la escala de mentira. Posteriormente de calcular los PB, se consultan las Normas del IAC para obtener los Puntajes T que es el puntaje que se utiliza para ubicar el nivel de cada área y del autoestima total [6].

4.3.3. Cuestionario de Hábitos de Estudio

Rondón [34] define hábitos de estudio como conductas que manifiesta el estudiante en forma regular ante el acto de estudiar y que repite constantemente.

Los hábitos de estudio se evalúan mediante el cuestionario elaborado por la Coordinación de Enseñanza de Programas Académicos de Enseñanza Media Superior, de la Universidad Nacional Autónoma de México. El Cuestionario de Hábitos de Estudio consta de 50 preguntas de opción múltiple, agrupadas en 6 áreas:

1. Estudio Independiente (11 preguntas).
2. Habilidades de Lectura (10 preguntas).
3. Administración de Tiempo (10 preguntas).
4. Concentración (5 preguntas).
5. Lugar de Estudio (4 preguntas).
6. Habilidades para Procesar la Información (10 preguntas).

Las posibles respuestas a las preguntas son: Nunca, Ocasionalmente, Algunas veces, Frecuentemente o Siempre. Los valores de las respuestas son 1, 2, 3, 4 y 5, respectivamente o 5, 4, 3, 2 y 1, dependiendo de la pregunta. Para obtener el puntaje de este cuestionario se suman los valores de todas las respuestas, esta suma se divide entre 50 y el resultado se multiplica por 2. El puntaje resultante esta entre 0 y 10 puntos.

4.3.4. Prueba de aula de razonamiento científico (Prueba de Lawson)

Una prueba que se utiliza para medir el razonamiento científico es la Prueba de aula de razonamiento científico o Prueba de Lawson. Esta prueba se diseñó para evaluar la capacidad de razonamiento científico de acuerdo a las propuestas de Piaget. La prueba consta de 12 preguntas que miden seis aspectos del razonamiento [30]:

1. Conservación de magnitudes física.
2. Pensamiento de proporcionalidad.
3. Identificación y control de variables.

4. Pensamiento probabilístico.

5. Pensamiento combinatorio.

6. Pensamiento correlacional.

Cada pregunta incluye un problema, del cual se responde la solución y el razonamiento que se usó para llegar a ésta, ambas son respuestas de opción múltiple. Para la calificación de esta prueba es necesario que los estudiantes seleccionen tanto la respuesta correcta como la razón correcta. Si la respuesta y la razón son correctas se concede un punto y si una o ambas de éstas son incorrectas, no se concede ninguna puntuación. Hay 12 preguntas en la prueba de Lawson, por lo que la puntuación es de 0 a 12 puntos [30].

De acuerdo con la cantidad de aciertos obtenidos, un estudiante se puede ubicar en uno de los siguientes niveles de razonamiento científico ([4], [11]):

- Empírico - Inductivo (Concreto, 0-4 puntos): Estudiantes que no son capaces de contrastar hipótesis involucrando agentes causales observables. Pueden llevar a cabo experimentos mentales. Las operaciones que usa son concretas, se relacionan directamente con objetos y no con hipótesis verbalizadas.

- Transición o Intermedio (Transición, 5-8 puntos): Para desarrollar este estado debe haber desarrollado previamente el pensamiento concreto. Estudiantes inconsistentemente capaces de contrastar hipótesis involucrando agentes observables causales. En este estado el individuo es capaz de razonar con proposiciones sin la necesidad de objetos, formular hipótesis y probarlas.

- Hipotético - Deductivo (Formal, 9-12 puntos): Estudiantes consistentemente capaces de contrastar hipótesis involucrando agentes causales observables o estudiantes capaces de contrastar hipótesis involucrando entidades que no están observando. Un pensador formal puede formular hipótesis y probarlas [1].

4.3.5. Test de habilidades Lecto-Comprensivas Básicas

El Test de habilidades Lecto-Comprensivas Básicas (THLB) es un instrumento que mide el nivel de compresión lectora. En específico, mide las siguientes habilidades de lectura comprensiva [32]:

[1] Descripción de los niveles de razonamiento científico dada por Ates y Cataloglu [4].

- Identificar las relaciones que guardan entre sí los diversos elementos de un texto.
- Leer una oración, párrafo o texto, reconociendo las afirmaciones y/o los sentidos implícitos que contenga.
- Especificar el sentido preciso de las palabras y expresiones dentro de un texto.
- Hacer inferencias sobre las informaciones de un texto.

Este test consta de tres textos de tipo: argumentativo, descriptivo y narrativo; y esta conformado por 45 preguntas de opción múltiple [32]. En la FCFM-BUAP se aplicó una parte del THLB, la cual consta de dos textos considerados los más familiares a los estudiantes mexicanos, por lo que se consideraron un total de 30 preguntas.

4.3.6. Cuestionario Honey-Alonso de Estilos de Aprendizaje

Keefe [20] define a los estilos de aprendizaje como sigue: “los estilos de aprendizaje son los rasgos cognitivos, afectivos y psicológicos que sirven como indicadores relativamente estables, de cómo los alumnos perciben, interaccionan y responden a sus ambientes de aprendizaje”. El modelo de Aprendizaje experiencial de Kolb es uno de los modelos más influyentes en este campo [13].

En el modelo de Kolb [23], la generación del conocimiento se define como un proceso de transformación de la experiencia percibida. Contempla dos dimensiones: percepción y procesamiento. La primera está vinculada a la captación y conceptualización de la experiencia, mientras que la segunda, se relaciona con el pensamiento y la comprobación de la información novedosa. Al interior de cada factor se encuentran dos procesos dialécticos o etapas que intervienen en toda instancia de aprendizaje, correspondiéndole a la dimensión perceptual los de experiencia concreta y conceptualización abstracta y a la procesual, la observación reflexiva y la experiencia activa ([13], [23]).

Kolb [23] describe cuatro estilos de aprendizaje que tienen lugar a partir de la interacción producida entre los cuatro procesos (experiencia concreta, conceptualización abstracta, observación reflexiva y experiencia activa). Las modalidades de aprendizaje propuestas caracterizan a las personas como: adaptadoras, convergentes, divergentes y asimiladoras (Figura 4.1).

Honey y Mumford [17] partieron de la teoría de Kolb y coinciden con la idea original sobre los cuatro procesos. No obstante, a diferencia de Kolb, homologan procesos a estilos y los renombran como: estilo activo (experiencia

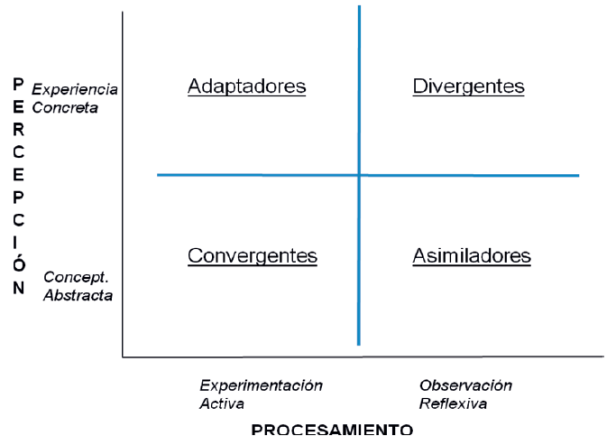

Figura 4.1: Modelo bidimensional de estilos de aprendizaje de Kolb. Fuente: [13].

concreta), estilo reflexivo (observación reflexiva), estilo teórico (conceptualización abstracta) y estilo pragmático (experiencia activa) [13].

Honey y Mumford [17] describen los estilos de aprendizaje a partir de las definiciones propuestas para los procesos por Kolb [23], manteniendo la esencia de sus ideas originales intacta. De acuerdo con esta reformulación teórica diseñan el *Learning Styles Questionnaire* (LSQ) para ser aplicado específicamente en el ámbito organizacional en el Reino Unido. Este cuestionario consta de 80 preguntas (20 referentes a cada estilo) de respuesta dicotómica (+ o -), a los que los examinados deben contestar según su acuerdo con cada afirmación. Del LSQ y de su modelo surge el Cuestionario Honey-Alonso de Estilos de Aprendizaje (CHAEA) [2], como la versión adaptada (métrica, conceptual y lingüísticamente) para el ámbito académico español. Respeta tanto la configuración teórica propuesta por Honey y Mumford sobre los cuatro estilos, como también la estructura de la herramienta original en cuanto al número de preguntas y a su modalidad de respuesta [13]. El cuestionario está diseñado para identificar el estilo preferido de aprender y la calificación es de 0 a 20 puntos para cada estilo (activo, reflexivo, teórico y pragmático).

Alonso, Gallego y Honey ([1], [2]) definen a los estilos de aprendizaje como:

- Activo: Corresponde a las personas que se caracterizan por ser animadoras, improvisadoras, descubridoras, espontáneas y arriesgadas. Están interesadas en vivir las experiencias y ser cambiantes.

- Reflexivo: Incluye a las personas que son ponderadas, receptivas, ana-

líticas y exhaustivas. Son observadoras, pacientes, detallistas, investigadoras y asimiladoras.

- Teórico: Caracteriza a las personas que son metódicas, lógicas, objetivas, críticas y estructuradas; son disciplinadas, ordenadas, buscadoras de hipótesis y teorías, además de exploradoras.
- Pragmático: Incluye a las personas experimentadoras, prácticas, eficaces y realistas; se caracterizan por ser rápidas, organizadoras, estar seguras de sí mismas, de solucionar problemas y de planificar sus acciones.

4.3.7. Cuestionario para alumnos de nuevo ingreso

El cuestionario para alumnos de nuevo ingreso fue elaborado en el proyecto Perfil Psicopedagógico de ingreso a las licenciaturas de la BUAP. El cuestionario consta de 24 preguntas (algunas de opción múltiple) con las cuales se miden características del alumno:

- Género.
- Estado civil.
- Características del bachillerato de procedencia.
- Elección de la carrera.
- Con quien vive.
- Recursos económicos.
- Materias reprobadas durante el bachillerato.

4.4. Análisis sin covariables

El análisis sin covariables del tiempo de deserción de los alumnos de las licenciaturas de la FCFM se realiza mediante modelos no paramétricos y paramétricos. Los modelos no paramétricos permiten hacer comparaciones de la estimación de la función de supervivencia y también de la estimación de la función de riesgo acumulado, por lo que mediante estos modelos se comparan las diferentes licenciaturas (LA, LF, LFA, LM y LMA) impartidas en la FCFM y también se comparan por género las licenciaturas en Actuaría, Matemáticas y Matemáticas Aplicadas (LA, LM y LMA).

El análisis paramétrico se realiza para las licenciaturas en Actuaría, Matemáticas y Matemáticas Aplicadas (LA, LM y LMA). Mediante este análisis se obtiene un modelo paramétrico para el tiempo de deserción de cada licenciatura, estos modelos permiten comparar las estimaciones de la función de supervivencia y las estimaciones de la función de riesgo de las tres licenciaturas analizadas.

4.4.1. Análisis no paramétrico

En el análisis no paramétrico, primero se estima la función de supervivencia de las diferentes licenciaturas (LA, LF, LFA, LM y LMA) de la FCFM con el estimador de Kaplan-Meier. En la Figura 4.2 se muestra el estimador de la función de supervivencia y el de la función de riesgo acumulado de la LA,

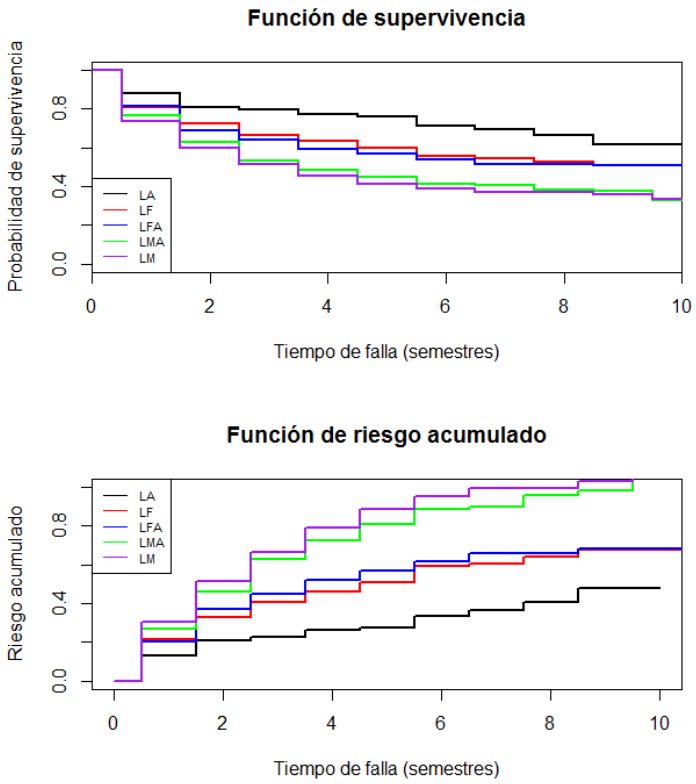

Figura 4.2: Estimación de la función de supervivencia y la función de riesgo acumulado de la LA, la LF, la LFA, la LM y la LMA, utilizando el estimador de Kaplan-Meier.

la LF, la LFA, la LM y la LMA, utilizando el estimador de Kaplan-Meier. Se observa que:

- La LA tiene mayor supervivencia estimada que la LF y la LFA. Además la LA, la LF y la LFA tienen mayor supervivencia estimada que la LM y la LMA (Figura 4.3). En este estudio se profundizará más en la LA, la LM y la LMA.
- Las estimaciones correspondientes a la LM y la LMA son similares, lo cual también ocurre con las de la LF y la LFA (Figura 4.5).
- En todas las licenciaturas (LA, LF, LFA, LM y LMA) de la FCFM, la función de riesgo acumulado estimada crece más en los primeros semes-

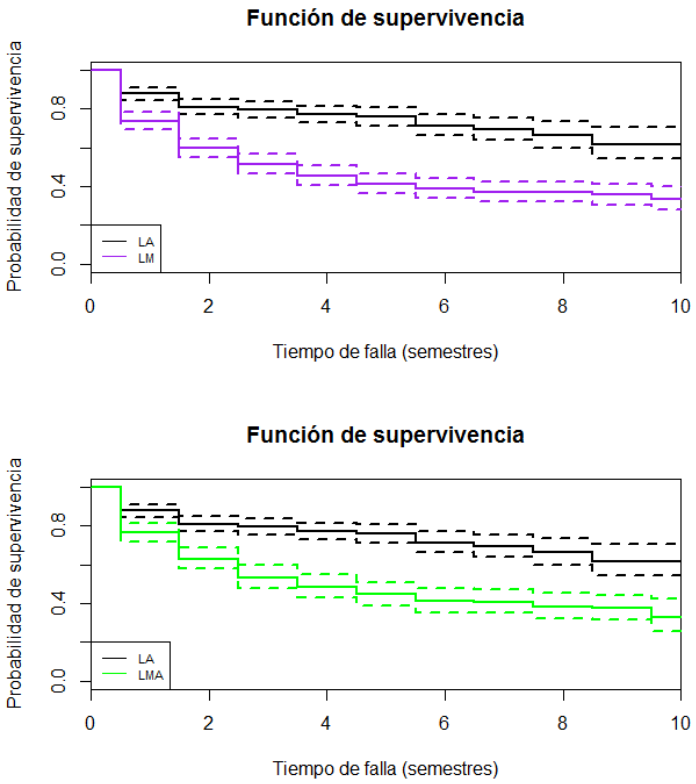

Figura 4.3: Comparación de las estimaciones de la función de supervivencia (con banda de confianza de 95 %) de la LA, la LM y la LMA, utilizando el estimador de Kaplan-Meier.

tres, lo cual indica que en los primeros semestres el riesgo de desertar es mayor.

En las Figuras 4.7 y 4.8 se muestra la estimación de la función de supervivencia y de la función de riesgo acumulado correspondientes a la LA, la LM y la LMA, con banda de confianza de 95 % y la estimación de la mediana. Se observa que:

- El comportamiento de la LM y la LMA es similar, se tiene que el riesgo acumulado aumenta más en el primer semestre, lo cual se debe al gran porcentaje de deserción del primer semestre (Figura 4.6).
- En el caso de la LA el riesgo acumulado también aumenta más en el primer semestre, pero en menor medida que en la LM y la LMA (Figura

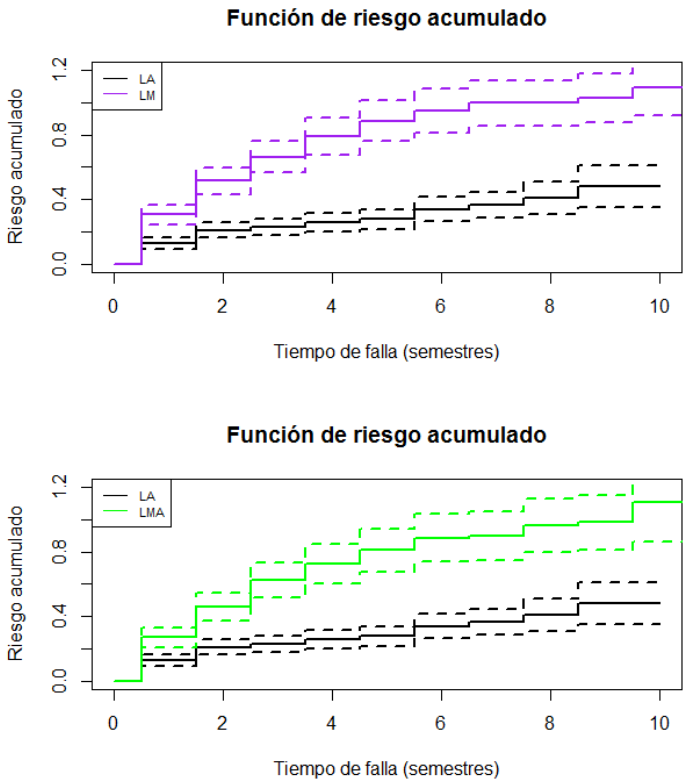

Figura 4.4: Comparación de las estimaciones de la función de riesgo acumulado (con banda de confianza de 95 %) de la LA, la LM y la LMA, utilizando el estimador de Kaplan-Meier.

4.4).

- Para la LM y la LMA, la estimación de la mediana es de 3 semestres y medio por lo que se puede decir que el 50 % de los estudiantes de las licenciaturas analizadas, desertan en los 3 primeros semestres y medio.

Las Figuras 4.9 y 4.10 muestran la comparación por género de la estimación de la función de supervivencia y de la función de riesgo acumulado de la LA, la LM y la LMA. Se observa que:

- En la LM no hay diferencia significativa entre las estimaciones de la función de supervivencia para cada género, lo cual también ocurre con las estimaciones de la función de riesgo acumulado.

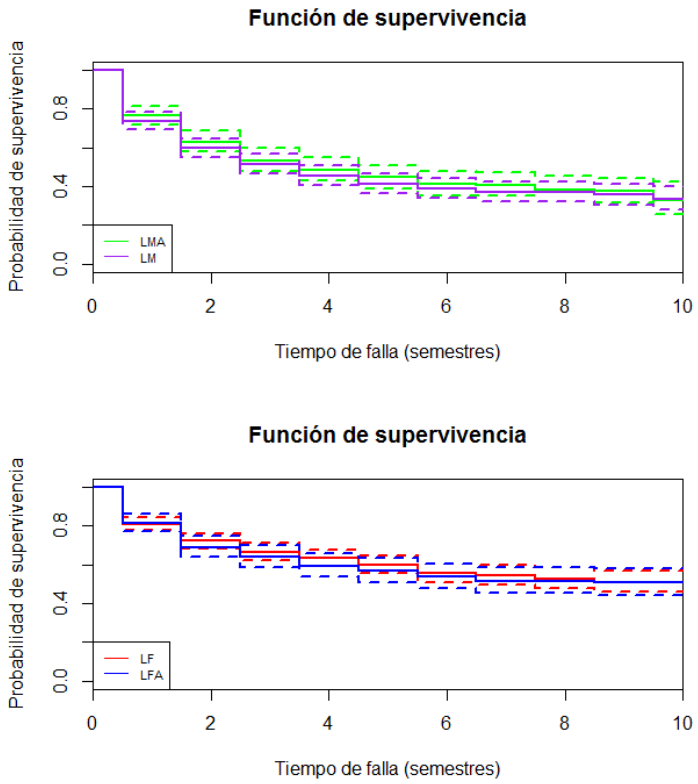

Figura 4.5: Comparación de las estimaciones de la función de supervivencia (con banda de confianza de 95 %) de la LM con la LMA y de la LF con la LFA, utilizando el estimador de Kaplan-Meier.

- La diferencia entre las estimaciones de la función de supervivencia para hombres y mujeres, así como entre las estimaciones de la función de riesgo acumulado, es mayor en la LMA que en la LA. En la LA tienen mayor supervivencia estimada las mujeres a partir del segundo semestre y medio y en la LMA tienen mayor supervivencia estimada los hombres a partir del tercer semestre y medio.

Resultados del análisis no paramétrico

El análisis no paramétrico muestra que el tiempo de deserción varía dependiendo de la licenciatura, los alumnos con mayor supervivencia son los

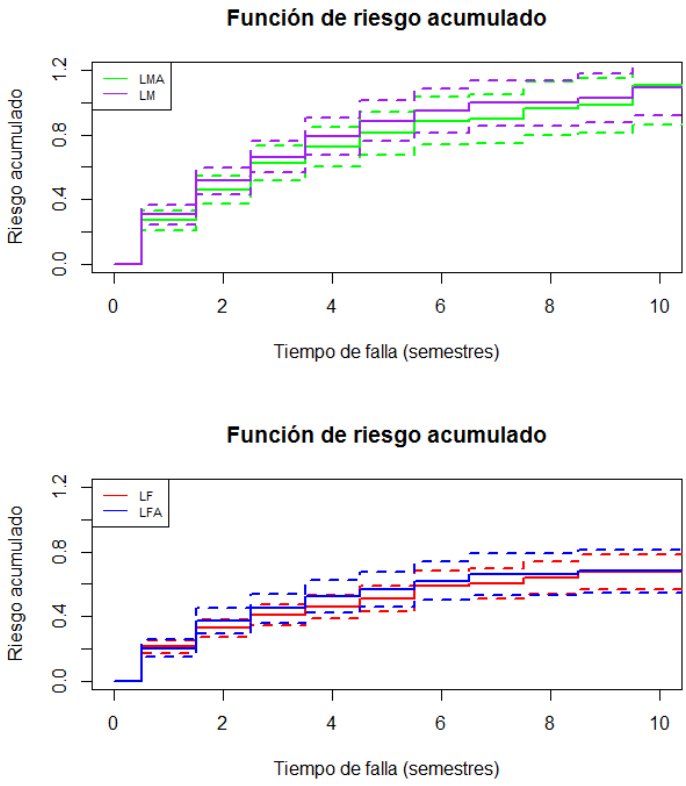

Figura 4.6: Comparación de las estimaciones de la función de riesgo acumulado (con banda de confianza de 95 %) de la LM con la LMA y de la LF con la LFA, utilizando el estimador de Kaplan-Meier.

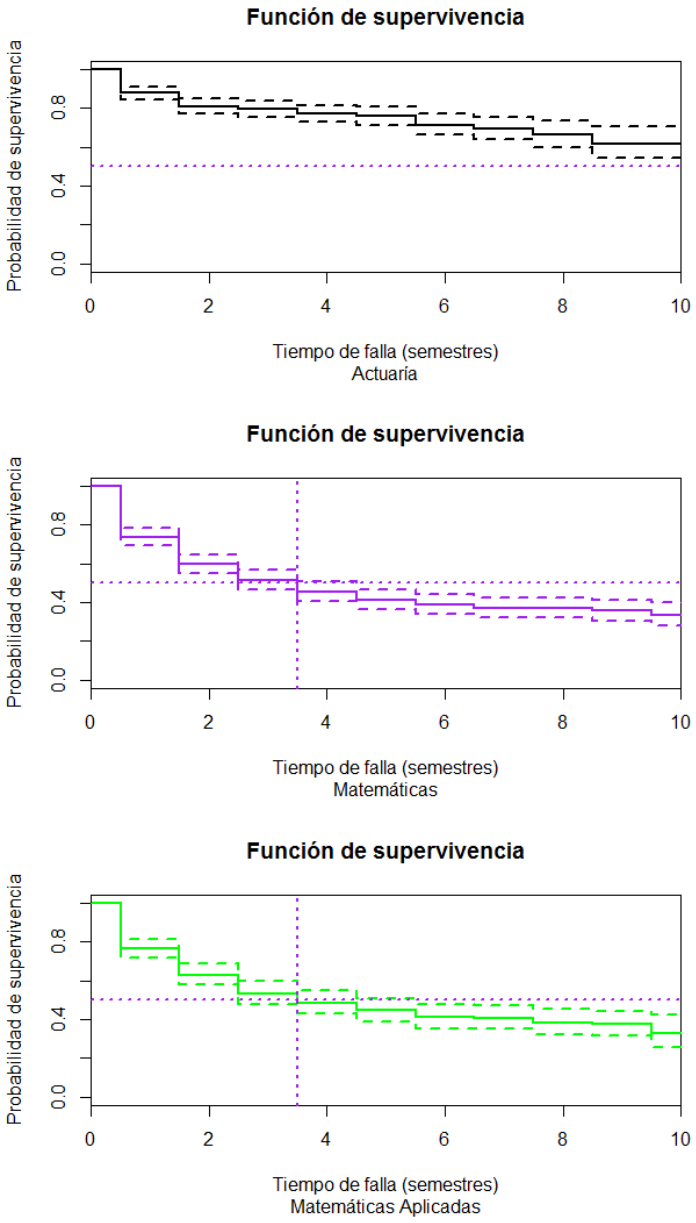

Figura 4.7: Estimación de la función de supervivencia (con banda de confianza de 95 % y estimación de la mediana) de la LA, la LM y la LMA, utilizando el estimador de Kaplan-Meier.

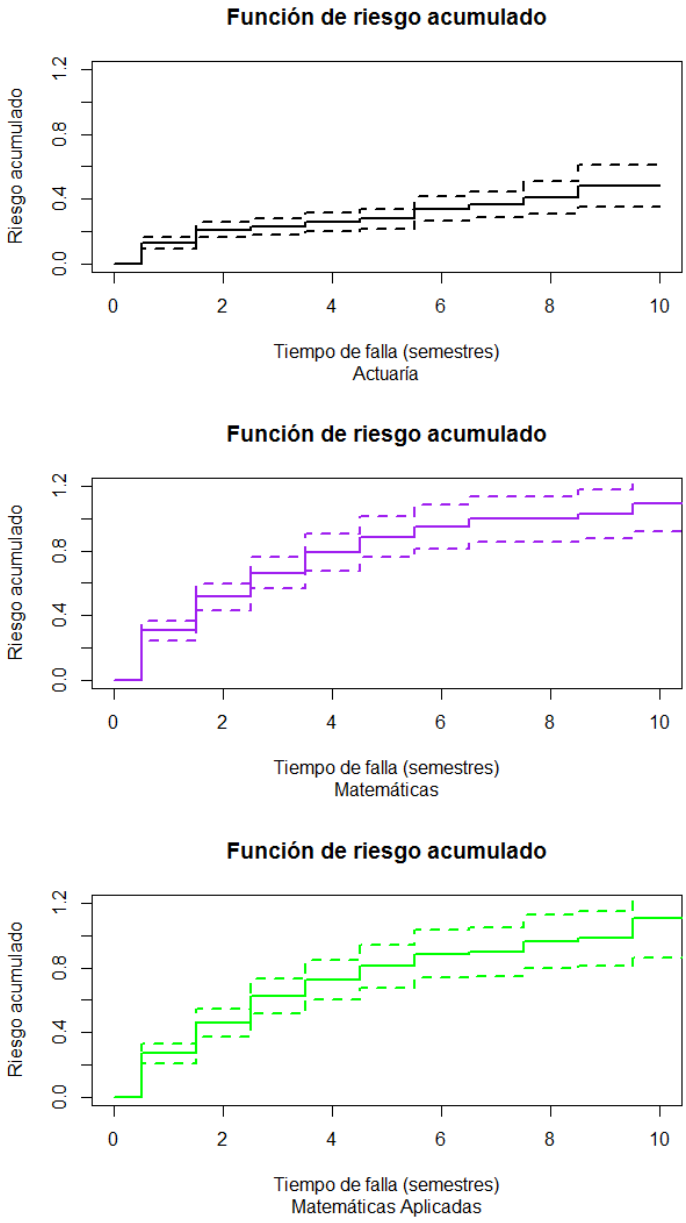

Figura 4.8: Estimación de la función de riesgo acumulado (con banda de confianza de 95 %) de la LA, la LM y la LMA, utilizando el estimador de Kaplan-Meier.

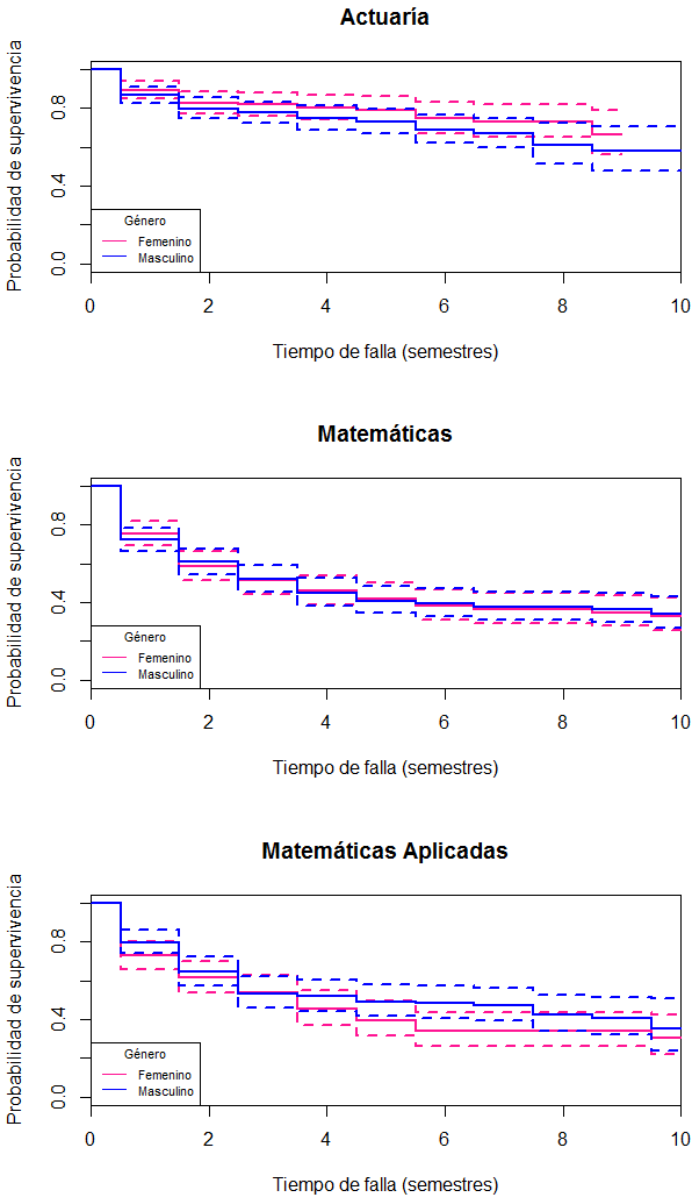

Figura 4.9: Comparación por género de la estimación de la función de supervivencia (con banda de confianza de 95 %) de la LA, la LM y la LMA, utilizando el estimador de Kaplan-Meier.

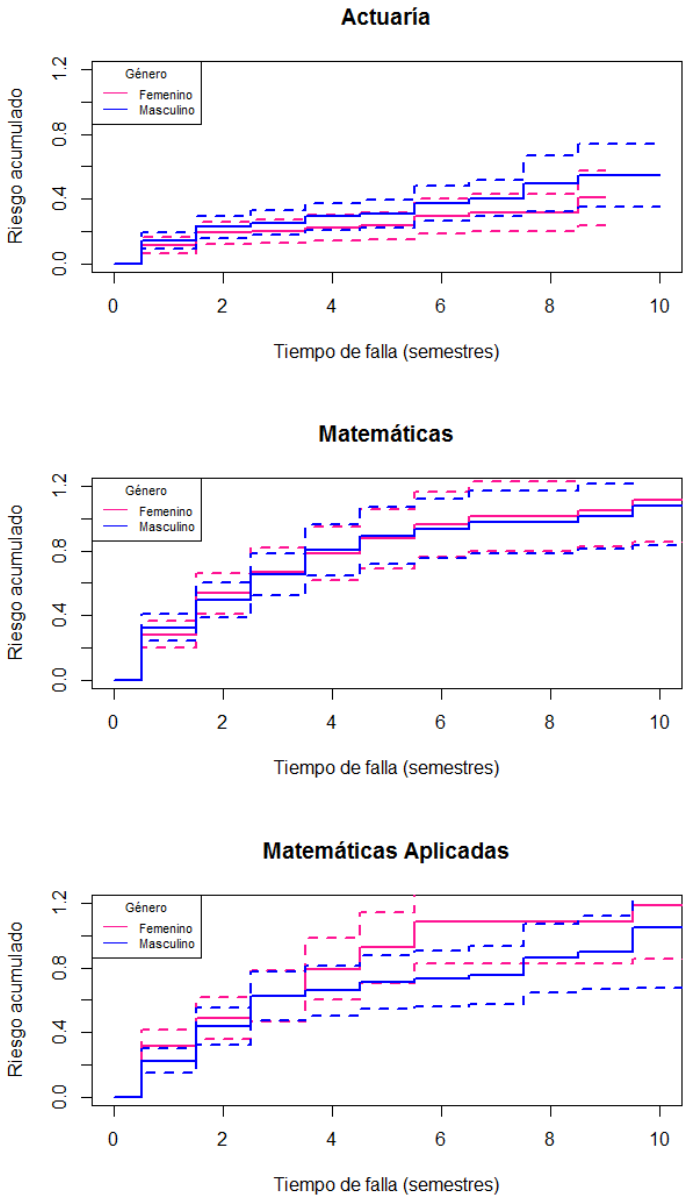

Figura 4.10: Comparación por género de la estimación de la función de riesgo acumulado (con banda de confianza de 95 %) de la LA, la LM y la LMA, utilizando el estimador de Kaplan-Meier.

Lic.	Exponencial	Weibull	$\log$ normal	Gama
LA	$Exp(0.063)$	$W(0.022, 0.532)$	$\log N(3.450, 8.756)$	$G(0.482, 0.010)$
LM	$Exp(0.163)$	$W(0.135, 0.558)$	$\log N(1.282, 4.392)$	$G(0.454, 0.047)$
LMA	$Exp(0.153)$	$W(0.126, 0.601)$	$\log N(1.403, 4.012)$	$G(0.504, 0.052)$

Tabla 4.5: Ajustes obtenidos para la LA, LM y LMA.

alumnos de la licenciatura en Actuaría y los alumnos con menor supervivencia son los de las licenciaturas de matemáticas (LM y LMA). También se observa que la supervivencia en la licenciatura en Matemáticas y en la licenciatura en Matemáticas Aplicadas es similar, lo cual también ocurre en las licenciaturas en Física y Física Aplicada.

En las diferentes licenciaturas la mayor deserción se da en los primeros semestres, sobre todo en el primero. En el caso de las licenciaturas de matemáticas (LM y LMA) la mayor deserción se da en los 4 primeros semestres.

4.4.2. Análisis paramétrico

En esta parte del análisis se supone que los tiempos de deserción (en la ausencia de censura) pueden ajustarse a alguna de las siguientes distribuciones: exponencial, Weibull, $\log$ normal o gama. La inferencia estadística realizada se basa en la metodología de máxima verosimilitud, en la Tabla 4.5 se muestran los ajustes que se obtienen mediante los estimadores de máxima verosimilitud para la LA, LM y LMA.

En la Figura 4.11 se comparan las estimaciones de la función de supervivencia que se obtienen mediante el estimador de Kaplan-Meier y mediante el análisis paramétrico. En las tres licenciaturas analizadas, se tiene que las estimaciones que proporcionan las distribuciones Weibull, $\log$ normal y gama son similares y ajustan mejor que la estimación que proporciona la distribución exponencial.

Los ajustes obtenidos en el análisis paramétrico proporcionan una estimación para la función de riesgo, en la Figura 4.12 se muestran las gráficas de las estimaciones de la función de riesgo para las diferentes licenciaturas analizadas. Se observa que:

- Las estimaciones de la función de riesgo proporcionadas por las distribuciones Weibull y gama son decrecientes aproximándose a cero, mientras que la estimación proporcionada por la distribución $\log$ normal es creciente y luego decreciente de manera similar a las estimaciones proporcionadas por las distribuciones Weibull y gama.

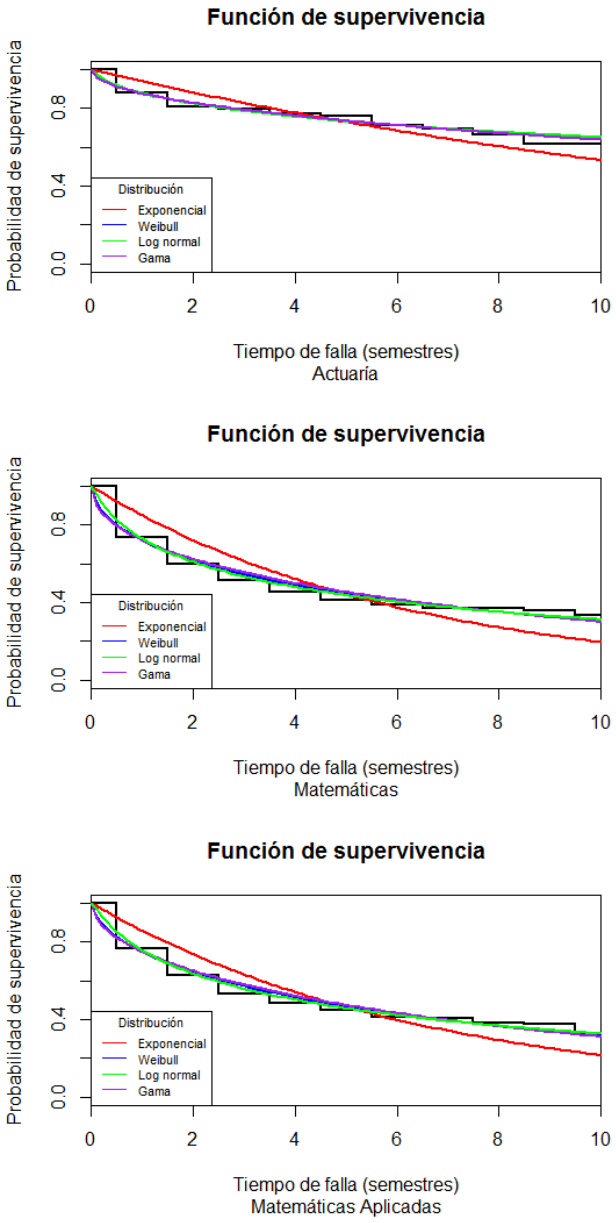

Figura 4.11: Estimaciones de la función de supervivencia mediante el estimador de Kaplan-Meier y cuando se supone que el tiempo de deserción tiene una distribución exponencial, Weibull, log normal o gama.

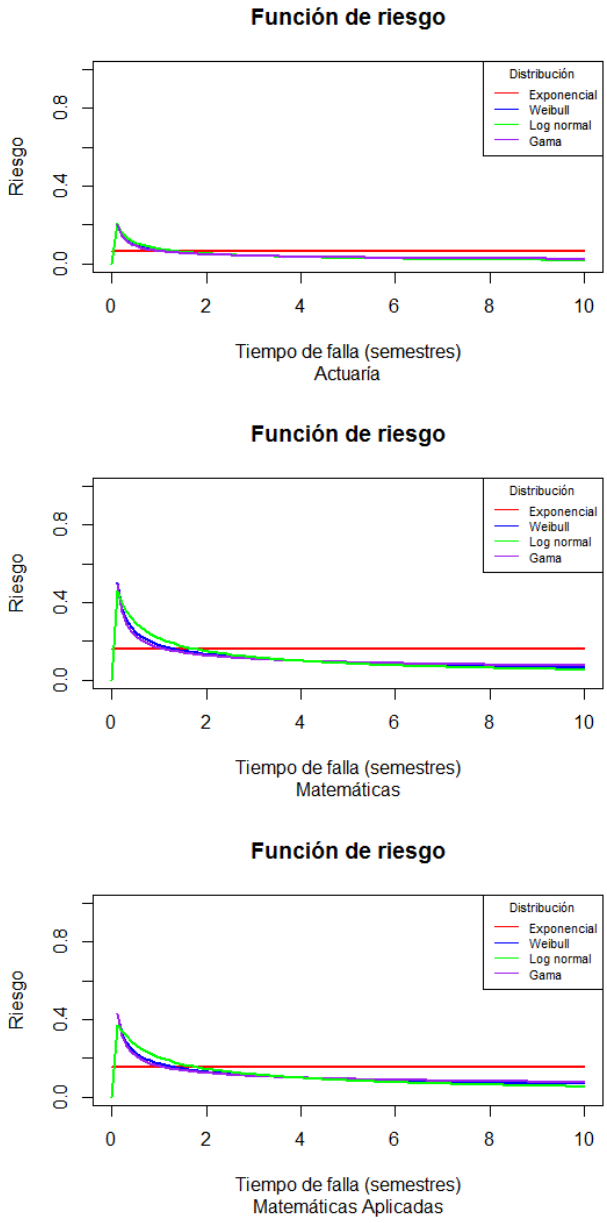

Figura 4.12: Estimación de la función de riesgo cuando se supone que el tiempo de deserción tiene una distribución exponencial, Weibull, log normal o gama.

Lic.	Exponencial		Weibull		$\log$ normal		Gama	
	CIA	CIB	CIA	CIB	CIA	CIB	CIA	CIB
LA	770.20	774.23	736.97	745.04	737.69	745.75	736.74	744.80
LM	1309.36	1313.38	1232.45	1240.49	1225.10	1233.13	1237.31	1245.35
LMA	986.89	990.65	943.71	951.23	939.14	946.66	946.35	953.87

Tabla 4.6: CIA y CIB de los diferentes modelos ajustados para la LA, LM y LMA.

- En las diferentes licenciaturas (LA, LM y LMA) se tiene que el riesgo de desertar es mayor en los primeros semestres, sobre todo en el primer semestre.

- Los alumnos de la LA tienen menor riesgo de desertar que los alumnos de la LM y la LMA.

Las Figuras 4.11 y 4.12 muestran que los ajustes obtenidos cuando se supone que el tiempo de deserción tiene una distribución Weibull, $\log$ normal o gama son similares, por lo que se utiliza el criterio de información de Akaike (CIA) y el criterio de información bayesiano (CIB) para seleccionar a uno de estos modelos. En la Tabla 4.6 se muestran los CIA y los CIB para los diferentes modelos ajustados. Comparando los CIA y los CIB, se tiene que en la licenciatura en Actuaría (LA) se elige el modelo de la distribución gama, mientras que en las licenciaturas de matemáticas (LM y LMA) se eligen los modelos de la distribución $\log$ normal.

En la Figura 4.13 se comparan las estimaciones que se eligieron mediante en CIA y el CIB para cada licenciatura, se observa que el comportamiento de las licenciaturas de matemáticas (LM y LMA) es similar y que los alumnos que tienen menor riesgo de desertar son los de Actuaría. También se nota que en las tres licenciaturas los alumnos tienen más riesgo de desertar en los primeros semestres, sobre todo en el primero.

Resultados del análisis paramétrico

En las licenciaturas en Actuaría, Matemáticas y Matemáticas Aplicadas (LA, LM y LMA), el análisis paramétrico proporciona tres modelos (Weibull, $\log$ normal y gama) que ajustan de manera similar a los datos. Comparando los CIA y los CIB de cada modelos ajustado, se tiene que en la licenciatura en Actuaría (LA) se prefiere el modelo de la distribución gama, mientras que en las licenciaturas de matemáticas (LM y LMA) se eligen los modelos de la

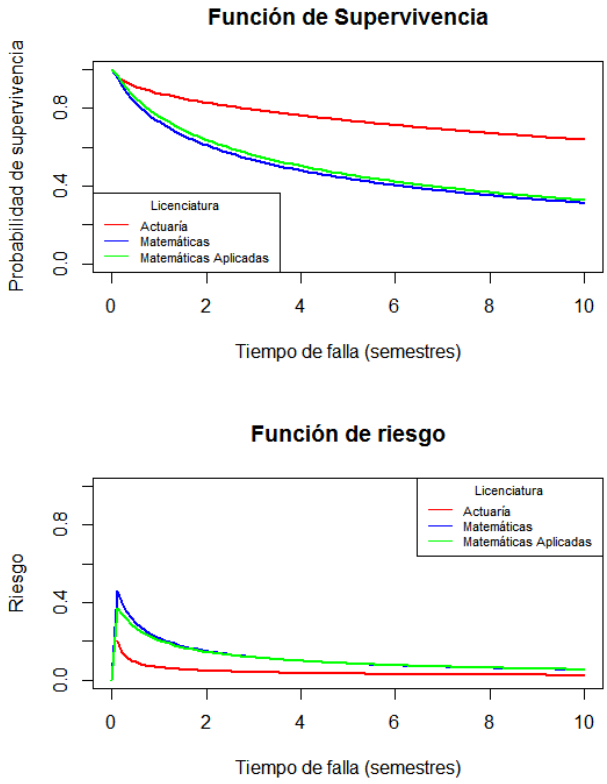

Figura 4.13: Comparación de las estimaciones de la función de supervivencia y la función de riesgo de las licenciaturas en Actuaría, Matemáticas y Matemáticas Aplicadas.

distribución $\log$ normal. Al comparar los modelos elegidos para cada licenciatura, nuevamente se observa que el comportamiento de las licenciaturas de matemáticas (LM y LMA) es similar y que los alumnos que tienen menor riesgo de desertar son los de Actuaría. También se nota que en las tres licenciaturas los alumnos tienen mayor riesgo de desertar en el primer semestre y después el riesgo es decreciente.

4.5. Análisis con covariables

El análisis con covariables del tiempo de deserción de los alumnos de las licenciaturas de la FCFM se realiza mediante el modelo de riesgo proporcional semiparamétrico, debido a que este modelo está enfocado en evaluar la

relación con los factores o covariables.

En el análisis con covariables primero se calcula la correlación de las covariables de escala mínima de intervalo. Posteriormente, se realiza el modelo de riesgo proporcional semiparamétrico con las covariables de escala mínima de intervalo y luego con todas las covariables consideradas en el estudio.

4.5.1. Correlación de las covariables

En el análisis se consideran las siguientes variables de escala mínima de intervalo: Puntaje, Autoestima, HabEstudio, Lawson, THLB, Activo, Reflexivo, Teórico, Pragmático. Al calcular la matriz de correlación de estas covariables se tiene que:

- En Actuaría (Tabla E.1), las covariables que tienen correlación positiva moderada ($0.4 \leqslant r < 0.7$) son:
  - HabEstudio y Teórico.
  - Reflexivo y Teórico.

  El resto de covariables tienen correlación positiva baja ($0.2 \leqslant r < 0.4$), positiva muy baja ($0 < r < 0.2$), nula ($r = 0$), negativa muy baja ($-0.2 < r < 0$) y negativa baja ($-0.4 < r \leqslant -0.2$).

- En Matemáticas (Tabla E.2), las covariables que tienen correlación positiva moderada son:
  - Puntaje y Lawson.
  - Puntaje y THLB.
  - HabEstudio y Teórico.
  - Reflexivo y Teórico.
  - Teórico y Pragmático.

  El resto de covariables tienen correlación positiva baja, positiva muy baja, nula, negativa muy baja y negativa baja.

- En Matemáticas Aplicadas (Tabla E.3), las covariables que tienen correlación positiva moderada son:
  - Puntaje y Lawson.
  - Reflexivo y Teórico.
  - Teórico y Pragmático.

El resto de covariables tienen correlación positiva baja, positiva muy baja, nula, negativa muy baja y negativa baja.

4.5.2. Modelo de riesgo proporcional semiparamétrico

En esta parte del análisis se evalúa la relación que tiene el tiempo de deserción con las covariables (factores causales) consideradas. Primero se realiza el modelo de riesgo proporcional semiparamétrico con las covariables de escala mínima de intervalo: Puntaje, Autoestima, HabEstudio, Lawson, THLB, Activo, Reflexivo, Teórico y Pragmático. Para las licenciaturas estudiadas (LA, LM y LMA), se obtienen modelos con la mayoría de covariables estadísticamente no significativas (Tablas E.4, E.5 y E.6).

Para obtener modelos con covariables estadísticamente significativas se propone realizar el procedimiento de regresión paso a paso (Apéndice C), que es una técnica de selección de modelo ampliamente utilizada en Regresión Lineal Múltiple. Esta técnica proporciona un modelo cuyas variables explicativas son estadísticamente significativas. El procedimiento que se realiza es el mismo que el que se lleva a cabo en Regresión Lineal Múltiple sólo que en este caso se estiman los coeficientes del modelo de riesgo proporcional en lugar de los de la regresión lineal múltiple, i.e. se consideran los p-valores de los estadísticos de las pruebas para la hipótesis $H_0 : \beta_k = 0$ correspondientes al modelo de riesgo proporcional en lugar de los p-valores correspondientes a la regresión lineal múltiple. En este análisis se considera $\alpha_{dentro} = 0.10 = \alpha_{fuera}$.

Para cada licenciatura (LA, LM y LMA), se obtiene un modelo de riesgo proporcional semiparamétrico mediante el procedimiento de regresión paso a paso. Estos modelos cumplen que todas sus covariables son estadísticamente significativas (Tablas E.7, E.8 y E.9). Posiblemente se podría disminuir el número de covariables significativas en el modelo si se disminuye el valor del nivel de significancia de entrada y salida. Los modelos se presentan a continuación.

- En Actuaría:

  $$\widehat{h}(t|\underline{x}) = h_0(t)\exp\{-0.118\text{Teórico} - 0.006\text{Puntaje} - 0.034\text{Autoestima}\},$$

  en donde $\underline{x} = (\text{Teórico}, \text{Puntaje}, \text{Autoestima})'$ y $h_0(\cdot)$ es la función de riesgo basal arbitraria.

- En Matemáticas:

  $$\widehat{h}(t|\underline{x}) = h_0(t)\exp\{-0.006\text{Puntaje}\},$$

Licenciatura	Variables significativas	$\exp(\widehat{\beta}_i)$	Efecto en $h(\cdot)$	
Actuaría	Estilo de aprendizaje Teórico (Teórico)	0.888	+	-
	Puntaje de ingreso (Puntaje)	0.994	+	-
	Autoestima (Autoestima)	0.967	+	-
Matemáticas	Puntaje de ingreso (Puntaje)	0.994	+	-
Matemáticas Aplicadas	Puntaje de ingreso (Puntaje)	0.993	+	-
	Razonamiento científico (Lawson)	1.102	-	+

Tabla 4.7: Interpretación del modelo de riesgo proporcional semiparamétrico con todas las covariables de escala mínima de intervalo para cada licenciatura. El Efecto en $h(\cdot)$ se indica como sigue: del lado izquierdo el efecto en el riesgo de una baja puntuación de la covariable, mientras que del lado derecho se indica el efecto de una puntuación alta. El simbolo + significa que el riesgo aumenta mientras que el simbolo - significa que el riesgo disminuye.

en donde $\underline{x} = (\text{Puntaje})$ y $h_0(\cdot)$ es la función de riesgo basal arbitraria.

- En Matemáticas Aplicadas:

$$\widehat{h}(t|\underline{x}) = h_0(t)\exp\{-0.008\text{Puntaje} + 0.097\text{Lawson}\},$$

en donde $\underline{x} = (\text{Puntaje}, \text{Lawson})'$ y $h_0(\cdot)$ es la función de riesgo basal arbitraria.

Estos modelos indican que la covariable Puntaje es estadísticamente significativa en las licenciaturas analizadas y que a mayor Puntaje es menor el riesgo de deserción. En el caso de Actuaría, otras covariables que con mayor puntuación indican menor riesgo de deserción son Teórico y Autoestima. Mientras que en Matemáticas Aplicadas parece que mayor puntuación en Lawson indica mayor riesgo de deserción, quizá esto se debe a la correlación positiva moderada existente entre Puntaje y Lawson. Estas interpretaciones se resumen en la Tabla 4.7. En las Figuras E.1 y E.2 se muestra la estimación de la función de supervivencia y de la función de riesgo acumulado, respectivamente, para los modelos de riesgo proporcional semiparamétrico obtenidos mediante el procedimiento de regresión paso a paso con todas las covariables de escala mínima de intervalo.

En la siguiente parte del análisis, para cada licenciatura estudiada se obtiene un modelo de riesgo proporcional semiparamétrico mediante el procedimiento de regresión paso a paso considerando a las covariables: Puntaje, Autoestima, HabEstudio, Lawson, THLB, Activo, Reflexivo, Teórico, Pragmático, Género, Vive, FinBach, TipoBach, MatRep, OpCarrera, SosEst, Trabajo y RecSem. En el análisis de Actuaría no se considera a la covariable OpCarrera ya que en esta licenciatura se llena el cupo, lo cual no permite que ingresen

alumnos que tenían como primera opción otra licenciatura de la BUAP. Los modelos obtenidos se presentan a continuación.

En Actuaría:

$$\widehat{h}(t|\underline{x}) = h_0(t)\exp\{0.535\text{MatRep} - 0.093\text{Teórico} - 0.164\text{Lawson} - 1.071\text{TipoBach}\},$$

en donde $\underline{x} = (\text{MatRep}, \text{Teórico}, \text{Lawson}, \text{TipoBach})'$ y $h_0(\cdot)$ es la función de riesgo basal arbitraria.

En Matemáticas:

$$\widehat{h}(t|\underline{x}) = h_0(t)\exp\{-0.006\text{Puntaje} + 0.846\text{Trabajo} + 0.717\text{OpCarrera} - 0.589\text{TipoBach}\},$$

en donde $\underline{x} = (\text{Puntaje}, \text{Trabajo}, \text{OpCarrera}, \text{TipoBach})'$ y $h_0(\cdot)$ es la función de riesgo basal arbitraria.

En Matemáticas Aplicadas:

$$\widehat{h}(t|\underline{x}) = h_0(t)\exp\{-0.007\text{Puntaje} - 0.560\text{RecSem} + 0.603\text{FinBach} + 0.095\text{Lawson}\},$$

en donde $\underline{x} = (\text{Puntaje}, \text{RecSem}, \text{FinBach}, \text{Lawson})'$ y $h_0(\cdot)$ es la función de riesgo basal arbitraria.

El modelo obtenido para Actuaría muestra que mayor puntuación en Teórico indica menor riesgo de deserción, lo cual también ocurre con la puntuación de Lawson. Se nota que los alumnos que proceden de un bachillerato general (TipoBach=0) tienen mayor riesgo de desertar que los que proceden de un bachillerato especializado (TipoBach=1). También se observa que a mayor número de materias reprobadas durante el bachillerato (MatRep) mayor es el riesgo de desertar.

El modelo para la Licenciatura en Matemáticas muestra que mayor Puntaje indica menor riesgo de deserción. También, los alumnos que proceden de un bachillerato general (TipoBach=0) tienen mayor riesgo de desertar que los que proceden de un bachillerato especializado (TipoBach=1). En esta licenciatura, los alumnos que trabajan (Trabajo=1) tienen mayor riesgo de desertar que los alumnos que no trabajan (Trabajo=0). Además, los alumnos que eligieron estudiar Matemáticas como primera opción (OpCarrera=0) tienen menor riesgo de desertar que los que la estudian como segunda opción (OpCarrera=1).

Licenciatura	Variables significativas	$\exp(\widehat{\beta}_i)$	Efecto en $h(\cdot)$	
Actuaría	Materias reprobadas (MatRep)	1.707	-	+
	Estilo de aprendizaje Teórico (Teórico)	0.912	+	-
	Razonamiento científico (Lawson)	0.849	+	-
	Tipo de bachillerato (TipoBach)	0.343	+	-
Matemáticas	Puntaje de ingreso (Puntaje)	0.994	+	-
	Trabajo (Trabajo)	2.331	-	+
	Opción de carrera (OpCarrera)	2.048	-	+
	Tipo de bachillerato (TipoBach)	0.555	+	-
Matemáticas Aplicadas	Puntaje de ingreso (Puntaje)	0.993	+	-
	Recursos semanales (RecSem)	0.571	+	-
	Financiamiento del bachillerato (FinBach)	1.827	-	+
	Razonamiento científico (Lawson)	1.100	-	+

Tabla 4.8: Interpretación del modelo de riesgo proporcional semiparamétrico con todas las covariables para cada licenciatura. El Efecto en $h(\cdot)$ se indica como sigue: del lado izquierdo el efecto en el riesgo de una baja puntuación de la covariable, mientras que del lado derecho se indica el efecto de una puntuación alta. El simbolo + significa que el riesgo aumenta mientras que el simbolo - significa que el riesgo disminuye.

El modelo para Matemáticas Aplicadas señala que mayor Puntaje indica menor riesgo de deserción y que mayor puntuación en Lawson indica mayor riesgo de deserción, esto posiblemente se debe a la correlación positiva moderada existente entre Puntaje y Lawson. Este modelo también revela que los alumnos con más recursos semanales (RecSem) tienen menor riesgo de desertar y que los alumnos que proceden de un bachillerato público (FinBach=0) tienen menor riesgo de deserción que los que proceden de un bachillerato privado (FinBach=1). Estas interpretaciones se resumen en la Tabla 4.8. En las Figuras E.3 y E.4 se muestra la estimación de la función de supervivencia y de la función de riesgo acumulado, respectivamente, para los modelos de riesgo proporcional semiparamétrico obtenidos mediante el procedimiento de regresión paso a paso con todas las covariables.

Dado que en las licenciaturas en matemáticas (Matemáticas y Matemáticas Aplicadas) la covariable Puntaje es indicadora de mayor riesgo de deserción, se realiza una comparación de las estimaciones de la función de supervivencia y de la función de riesgo acumulado para los alumnos con Puntaje mayor o igual a 750 con los alumnos con Puntaje menor a 750. La Figura 4.14 muestra que los alumnos con Puntaje mayor o igual a 750 tienen mayor supervivencia estimada que los alumnos con Puntaje menor a 750. Además, la Figura 4.15 muestra que los alumnos con Puntaje mayor o igual a 750 ya no están en riesgo de desertar a partir del quinto semestre y medio en la LM

y a partir del séptimo semestre y medio en la LMA.

Resultados del análisis con covariables

Para cada licenciatura estudiada se obtienen dos modelos de riesgo proporcional semiparamétricos, uno considerando sólo covariables de escala mínima de intervalo y el otro considerando a todas las covariables. Estos modelos sólo contienen covariables estadísticamente significativas y muestran que valores en las covariables indican mayor riesgo de deserción.

Los modelos obtenidos cuando sólo se consideran a las covariables de escala mínima de intervalo muestran que el Puntaje es una covariable significativa en las licenciaturas estudiadas y que mayor Puntaje indica menor riesgo de deserción. En la Licenciatura en Actuaría otras covariables que tie-

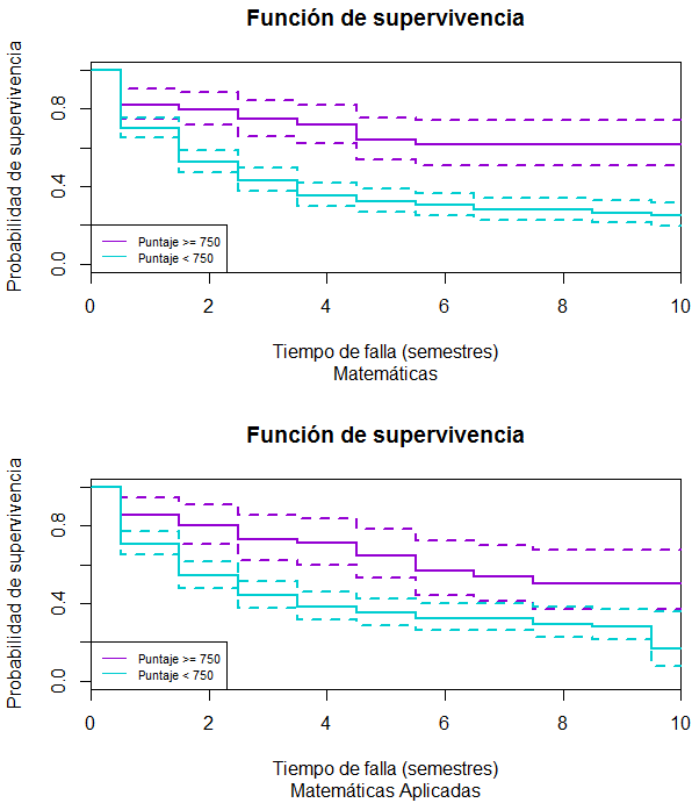

Figura 4.14: Comparación de las estimaciones de la función de supervivencia por Puntaje (con banda de confianza de 95 %) de la LM y la LMA, utilizando el estimador de Kaplan-Meier.

nen la misma interpretación son Teórico y Autoestima.

Los modelos resultantes cuando se consideran a todas las covariables incluyen al menos una covariable académica (Teórico, Lawson o Puntaje) y una covariable que tiene que ver con las características del bachillerato de procedencia (TipoBach o FinBach). Esto significa que el bachillerato de procedencia influye en el riesgo de deserción de los alumnos de las licenciaturas estudiadas.

Sólo en las licenciaturas de matemáticas (Matemáticas y Matemáticas Aplicadas), los modelos contienen una covariable que tiene que ver con la economía (Trabajo o RecSem). Lo cual se puede deber a la diferencia económica existente entre los alumnos de estas carreras y los de Actuaría.

Los modelos muestran que el número de materias reprobadas en el ba-

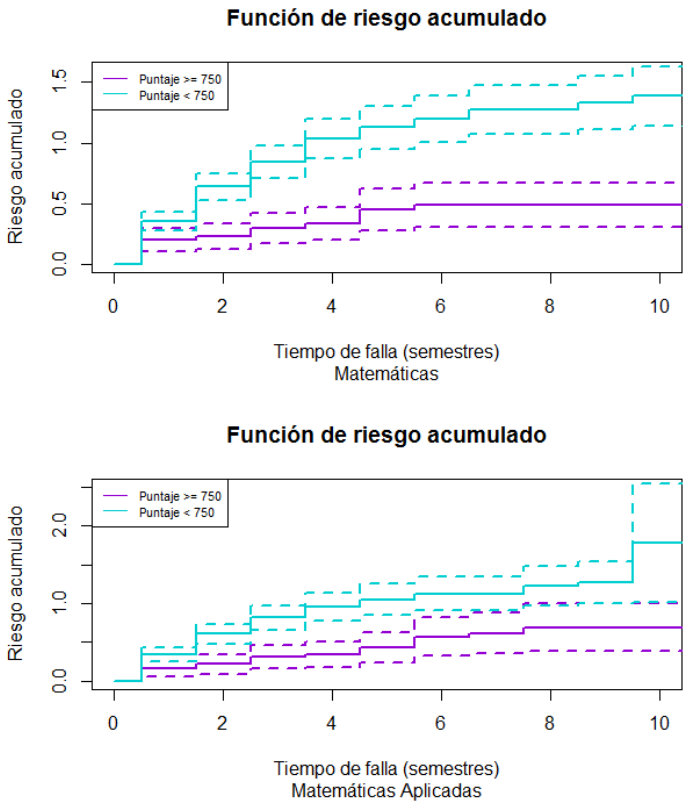

Figura 4.15: Comparación de las estimaciones de la función de riesgo acumulado por Puntaje (con banda de confianza de 95 %) de la LM y la LMA, utilizando el estimador de Kaplan-Meier.

chillerato (MatRep) sólo aumenta el riesgo de deserción de los estudiantes de Actuaría. Mientras que el hecho de estar estudiando la carrera como segunda opción (OpCarrera) sólo aumenta el riesgo de los estudiantes de Matemáticas.

La comparación por Puntaje muestra que en las licenciaturas en matemáticas (Matemáticas y Matemáticas Aplicadas) los alumnos con Puntaje mayor o igual a 750 tienen mayor supervivencia estimada que los alumnos con Puntaje menor a 750.

A pesar de que los modelos obtenidos varían, se puede decir que en las licenciaturas estudiadas el Puntaje de ingreso a la universidad es un factor que tiene gran influencia en la deserción. Además, el bachillerato de procedencia también influye en la deserción. También, características respecto a la economía sólo afectan a los estudiantes de las licenciaturas de matemáticas (Matemáticas y Matemáticas Aplicadas).

4.6. Conclusiones del análisis del caso de estudio

El análisis de la deserción en las licenciaturas de la FCFM-BUAP muestra que la deserción varía dependiendo de la licenciatura. De las cinco licenciaturas impartidas en la FCFM, los alumnos con mayor supervivencia son los alumnos de la Licenciatura en Actuaría y los alumnos con menor supervivencia son los de las licenciaturas en matemáticas (Matemáticas y Matemáticas Aplicadas). La supervivencia de los alumnos de la licenciatura en Matemáticas y los de la licenciatura en Matemáticas Aplicadas es similar, lo cual también ocurre con los alumnos de las licenciaturas en Física y Física Aplicada.

En las licenciaturas estudiadas (Actuaría, Matemáticas y Matemáticas Aplicadas) el semestre con más riesgo de desertar es el primero, en los semestres posteriores el riesgo decrece. Lo cual confirma lo que ya se ha reportado en la literatura, que en el primer año se dan los porcentajes más altos de deserción.

El análisis de deserción con covariables proporciona modelos de riesgo proporcional semiparamétrico para cada licenciatura bajo estudio. Cuando sólo se consideran covariables de escala mínima de intervalo, los modelos muestran que el puntaje de ingreso a la universidad (Puntaje) es un factor indicador de la decisión de desertar de los alumnos de las carreras consideradas.

Los modelos resultantes cuando se consideran a todas las covariables, muestran que la puntuación alta de una covariable académica (en Actuaría: Teórico o Lawson; en Matemáticas y Matemáticas Aplicadas: Puntaje) indica menor riesgo de desertar y que las características del bachillerato de proce-

dencia también influyen en el riesgo de deserción. Los alumnos de las licenciaturas en matemáticas (Matemáticas y Matemáticas Aplicadas) con Puntaje menor a 750 tienen más riesgo de desertar que los alumnos con Puntaje mayor o igual a 750.

Sólo los modelos correspondientes a las licenciaturas de matemáticas (Matemáticas y Matemáticas Aplicadas) indican que una covariable de economía esta relacionada con el riesgo de deserción. Esta es una diferencia que se debe considerar al crear estrategias de retención, ya que las causas de deserción para los alumnos de Actuaría no son las mismas que las de los alumnos de las licenciaturas de matemáticas (Matemáticas y Matemáticas Aplicadas).

Otra diferencia en los indicadores de riesgo de deserción es que el número de materias reprobadas en el bachillerato (MatRep) sólo afecta a los estudiantes de Actuaría. Mientras que el hecho de estar estudiando la carrera como segunda opción (OpCarrera) sólo afecta a los estudiantes de Matemáticas.

En la obtención de los modelos de riegos proporcional se consideran varias covariables, sin embargo sólo resultaron significativas algunas de ellas. Esto no significa que la deserción sólo se ve influenciada por esos factores, sino que en este análisis esos factores resultaron indicadores de mayor riesgo de deserción.

Conclusiones

El Análisis de Supervivencia consiste en un conjunto de técnicas que estudian el tiempo hasta que ocurre un evento específico. Al evento específico frecuentemente se le llama falla y el periodo de tiempo en el que ocurre es llamado el tiempo de falla. Su metodología es muy extensa ya que incluye una gran variedad de modelos (paramétricos, no paramétricos y semiparamétricos), además algunos modelos permiten incluir variables explicativas o covariables para evaluar la relación existente entre el tiempo de falla y las covariables. Los diferentes modelos de supervivencia, con o sin covariables, se pueden usar para estudiar diferentes datos, incluso cuando se tienen individuos de los que no se conoce el tiempo de falla exacto (censurados).

La flexibilidad del Análisis de Supervivencia permite aplicarlo en una variedad de estudios. En este trabajo se aplica esta metodología en un estudio de la deserción de las licenciaturas de la Facultad de Ciencias Físico Matemáticas (FCFM) de la Benemérita Universidad Autónoma de Puebla (BUAP). Se aplican diferentes modelos en el estudio de la deserción, los cuales proporcionan información de interés en el desarrollo de estrategias de retención estudiantil apropiadas para los estudiantes de esta facultad.

El análisis paramétrico y no paramétrico muestra que el tiempo de deserción varía dependiendo de la licenciatura, los alumnos con mayor supervivencia son los alumnos de la licenciatura en Actuaría y los alumnos con menor supervivencia son los de las licenciaturas de matemáticas (Matemáticas y Matemáticas Aplicadas). También se observa que la supervivencia en la licenciatura en Matemáticas y en la licenciatura en Matemáticas Aplicadas es similar, lo cual también ocurre en las licenciaturas en Física y Física Aplicada. En el caso de las licenciaturas de matemáticas (Matemáticas y Matemáticas Aplicadas) la mayor deserción se da en los 4 primeros semestres.

En las licenciaturas en Actuaría, Matemáticas y Matemáticas Aplicadas, el análisis proporciona un modelo paramétrico para cada licenciatura, estos modelos muestran que los alumnos que tienen menor riesgo de desertar en la FCFM son los de Actuaría y también que en las tres licenciaturas los alumnos tienen mayor riesgo de desertar en el primer semestre y después el riesgo es decreciente.

Para las licenciaturas en Actuaría, Matemáticas y Matemáticas Aplicadas el análisis de deserción con covariables proporciona modelos de riesgo proporcional semiparamétricos. Estos modelos muestran que mayor puntaje de

ingreso a la universidad (Puntaje) indica menor riesgo de desertar. Además, al considerar todas las covariables estudiadas los modelos resultantes muestran que la puntuación alta de una covariable académica (en Actuaría: Teórico o Lawson; en Matemáticas y Matemáticas Aplicadas: Puntaje) indica menor riesgo de desertar, que las características del bachillerato de procedencia influyen en el riesgo de deserción y que sólo los modelos correspondientes a las licenciaturas de matemáticas (Matemáticas y Matemáticas Aplicadas) indican que una covariable de economía (Trabajo o RecSem) esta relacionada con el riesgo de deserción.

Una opción para continuar con el estudio de la deserción en las licenciaturas de la FCFM-BUAP es estimar a la función de riesgo basal de los modelos de riesgo proporcional semiparamétricos para tener un modelo que permita hacer predicción o pronósticos. Otra opción es realizar el análisis a una base de datos en la que se incluya información de generaciones más recientes, de las cuales se midan las covariables consideradas en este estudio, sobre todo las que resultaron significativas en los modelos, y otros factores como los que se presentan en la Tabla 4.1, algunos factores de interés son: promedio en la licenciatura, materias reprobadas en la licenciatura, educación de los padres, relación con los profesores y relación con los compañeros, este último factor ha sido considerado predictor de la persistencia en cursos de física introductoria [44]. También, es de interés considerar en el análisis que algunas covariables dependen del tiempo, por ejemplo: trabajo, promedio de la licenciatura, materias reprobadas en la licenciatura. El análisis con covariables dependientes del tiempo se puede realizar a través de un modelo de regresión tipo Cox con puntos de cambio en las covariables, este modelo es de interés en el desarrollo de la teoría del Análisis de Supervivencia, por lo que es otra opción para un trabajo posterior.

Apéndice A

Conceptos estadísticos

En este Apéndice se muestran algunos resultados estadísticos con los cuales se desarrolla el Capítulo 2.

Teorema A.1. *Si el t-ésimo momento de una variable aleatoria continua X existe, entonces [39]*

$$\lim_{n\to\infty} n^t Pr(|X| > n) = 0.$$

Demostración. Dado que el t-ésimo momento de X existe

$$\infty > \int_{\mathbb{R}} |x|^t f(x)dx = \lim_{n\to\infty} \int_{|x|\leqslant n} |x|^t f(x)dx,$$

luego

$$\lim_{n\to\infty} \int_{|x|>n} |x|^t f(x)dx = 0.$$

Además,

$$\begin{aligned} \lim_{n\to\infty} \int_{|x|>n} |x|^t f(x)dx &\geqslant \lim_{n\to\infty} n^t \int_{|x|>n} f(x)dx \\ &= \lim_{n\to\infty} n^t Pr(|X| > n) \\ &\geqslant 0, \end{aligned}$$

por lo tanto

$$\lim_{n\to\infty} n^t Pr(|X| > n) = 0.$$

□

El reciproco no es necesariamente cierto, i.e. si

$$\lim_{n\to\infty} n^t Pr(|X| > n) = 0,$$

entonces el t-ésimo momento de una variable aleatoria continua X no necesariamente existe [39].

Lema A.2. *Si X es una variable aleatoria continua no negativa con f.d.a. $F(\cdot)$, entonces*

$$E(X) = \int_0^\infty (1 - F(x))dx, \tag{A.1}$$

si existe alguno de ambos lados de (A.1) [39].

Demostración. Para probar que la existencia del lado izquierdo de (A.1) implica que el lado derecho es finito y ambos lados son iguales, se asume que $E(X)$ existe, esto es

$$\infty > E(X) = \int_0^\infty xf(x)dx = \lim_{n\to\infty} \int_0^n xf(x)dx.$$

Integrando por partes se tiene

$$\begin{aligned}
\int_0^n xf(x)dx &= xF(x)|_0^n - \int_0^n F(x)dx \\
&= nF(n) - n + n - \int_0^n F(x)dx \\
&= n(F(n) - 1) + \int_0^n (1 - F(x))dx \\
&= -n(1 - F(n)) + \int_0^n (1 - F(x))dx \\
&= -nPr(X > n) + \int_0^n (1 - F(x))dx.
\end{aligned}$$

Como X es una v. a. no negativa

$$\int_0^n xf(x)dx = -nPr(|X| > n) + \int_0^n (1 - F(x))dx,$$

luego por el Teorema A.1 se tiene

$$\begin{aligned}
E(X) &= \lim_{n\to\infty} \left\{ -nPr(|X| > n) + \int_0^n (1 - F(x))dx \right\} \\
&= \lim_{n\to\infty} \int_0^n (1 - F(x))dx \\
&= \int_0^\infty (1 - F(x))dx.
\end{aligned}$$

Así, la existencia de $E(X)$ implica que $\int_0^\infty (1 - F(x))dx$ es finita y que ambos son iguales.

Ahora se mostrará que si $\int_0^\infty (1 - F(x))dx$ es finita, entonces $E(X)$ existe, i.e. $E(|X|) = E(X) < \infty$ y ambas expresiones son iguales. Dado que X es una v. a. no negativa

$$\begin{aligned}\int_0^n |x|f(x)dx &= \int_0^n xf(x)dx \\ &= -n(1 - F(n)) + \int_0^n (1 - F(x))dx,\end{aligned}$$

como se vio arriba. Dado que $-n(1 - F(n)) \leqslant 0$

$$\int_0^n |x|f(x)dx \leqslant \int_0^n (1 - F(x))dx \leqslant \int_0^\infty (1 - F(x))dx, \forall n.$$

Así

$$E(|X|) = \lim_{n\to\infty} \int_0^n |x|f(x)dx \leqslant \int_0^\infty (1 - F(x))dx < \infty.$$

Por lo tanto $E(X)$ existe y es igual a $\int_0^\infty 1 - F(x)dx$ como se mostró arriba. □

El siguiente resultado se usa frecuentemente en el desarrollo de procedimientos de inferencia para muestras grandes. Aquí "$\xrightarrow{D}$" significa "converge en distribución a".

Teorema A.3. *Sean $T_{1n}, \ldots, T_{kn}$ estadísticos para $(\theta_1, \ldots, \theta_k)$ tales que cuando $n \to \infty$*

$$\sqrt{n}(T_{1n} - \theta_1, \ldots, T_{kn} - \theta_k) \xrightarrow{D} N(\underline{0}, \Sigma)$$

donde $\Sigma = (\sigma_{ij})_{k\times k}$. Si $g(x_1, \ldots, x_k)$ es una función que tiene todas sus primeras derivadas, entonces cuando $n \to \infty$

$$\sqrt{n}(g(T_{1n}, \ldots, T_{kn}) - g(\theta_1, \ldots, \theta_k)) \xrightarrow{D} N\left(\underline{0}, \sum_{i=1}^k \sum_{j=1}^k \sigma_{ij} \frac{\partial g}{\partial \theta_i} \frac{\partial g}{\partial \theta_j}\right)$$

donde $\dfrac{\partial g}{\partial \theta_i}$ significa $\dfrac{\partial g(\theta_1, \ldots, \theta_k)}{\partial \theta_i}$, $i = 1, \ldots, k$ [25].

Un caso especial importante del Teorema A.3 ocurre cuando $k = 1$: si $\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2)$ cuando $n \to \infty$, entonces si $g(\cdot)$ tiene primera derivada $g'(\cdot)$,

$$\sqrt{n}[g(T_n) - g(\theta)] \xrightarrow{D} N(0, g'(\theta)^2\sigma^2). \tag{A.2}$$

Esto implica que, en las distribuciones asintóticas [25]

$$Var(g(T_n)) = g'(\theta)^2 Var(T_n). \tag{A.3}$$

Apéndice B

La integral producto

La integración del producto fue introducida por el matemático italiano Vito Volterra, como una herramienta en la solución de una cierta clase de ecuaciones diferenciales. Las ideas de integración de productos tienen una aparición muy natural en el Análisis de Supervivencia, y el desarrollo de este tema (en particular, del estimador de Kaplan-Meier) podría haber sido mucho más suave si la integración del producto hubiera sido un tema familiar desde el principio. El estimador de Kaplan-Meier es la integral producto del estimador de Nelson-Aalen de la función de riesgo acumulado; estos dos estimadores tienen la misma relación que hay entre la función de supervivencia y la función de riesgo acumulado. Existen muchas otras aplicaciones de la integración de productos en el Análisis de Supervivencia, por ejemplo en el estudio de procesos multiestados (conectados a la teoría de procesos de Markov) y en la teoría de la verosimilitud parcial [14].

Suponga que $X(\cdot)$ es una función de valor matricial ($p \times p$) del tiempo t. Suponga también que $X(\cdot)$ (o cada componente de $X(\cdot)$) es continua por la derecha con límite por la izquierda. Se denota a la matriz de identidad por I. La integral producto de $X(\cdot)$ sobre un intervalo $[0, t]$ se define como

$$\prod_0^t (I + dX(s)) = \lim_{\text{máx}|t_i - t_{i-1}| \to 0} \prod (I + (X(t_i) - X(t_{i-1})))$$

donde el límite se toma sobre una secuencia de particiones cada vez más finas $0 = t_0 < t_1 < \ldots < t_k = t$ del intervalo de tiempo $[0, t]$. Para que el límite exista, $X(\cdot)$ tiene que ser de variación acotada; equivalentemente, cada componente de $X(\cdot)$ es la diferencia de dos funciones crecientes [14].

Una propiedad de la integración del producto es su multiplicatividad. Definiendo la integral producto sobre un intervalo de tiempo arbitrario de la manera

natural se tiene que para $0 < s < t$ [14]

$$\prod_{0}^{t}(I + dX(s)) = \prod_{0}^{s}(I + dX(s)) \prod_{s}^{t}(I + dX(s)). \tag{B.1}$$

Si $G(\cdot)$ es continua en $(a, b]$ y $g(u) = G'(u)$, entonces la integral producto de la función $G(\cdot)$ cumple

$$\prod_{(a,b]} (1 + dG(s)) = \prod_{(a,b]} (1 + g(s)ds) = \exp\left\{\int_a^b g(u)du\right\} \tag{B.2}$$

que relaciona la integral producto y la integral de Riemann [25].

Apéndice C

Regresión paso a paso

La regresión paso a paso (*stepwise*) es quizás la técnica de selección de modelos más ampliamente usada en la regresión lineal múltiple. Se puede utilizar en situaciones donde hay un número muy grande de candidatas a ser variables independientes. La versión de la regresión paso a paso que se describirá está basada en los p-valores de los estadísticos de las pruebas para la hipótesis $H_0 : \beta_k = 0$. Antes de operar el algoritmo, el usuario elige dos p-valores de umbral, α_{dentro} y α_{fuera} con $\alpha_{dentro} \leqslant \alpha_{fuera}$. Los pasos para la regresión paso a paso son:

1. **Selección hacia adelante:** Se selecciona la variable independiente con el p-valor más pequeño, suponiendo que se satisface $p < \alpha_{dentro}$. Esta variable se introduce en el modelo, creando un modelo de una sola variable independiente.

2. También en un paso de selección hacia adelante, se revisan una a una las variables restantes como candidatas para la segunda variable en el modelo. La que tenga el p-valor más pequeño se agrega al modelo, suponiendo nuevamente que $p < \alpha_{dentro}$.

3. **Eliminación hacia atrás:** Es posible que al haber agregado la segunda variable al modelo se provoque un aumento en el p-valor de la primera variable. La primera variable se elimina del modelo si su p-valor es mayor que α_{fuera}.

4. El algoritmo continúa alternando los pasos de selección hacia adelante con los de eliminación hacia atrás: en cada paso de selección hacia adelante se agrega la variable con el p-valor más pequeño si $p < \alpha_{dentro}$, y en cada paso de eliminación hacia atrás se elimina la variable

con el p-valor más grande si $p > \alpha_{fuera}$. El algoritmo se termina cuando ninguna variable satisface los criterios para ser agregada o eliminada del modelo.

Una debilidad de todos los procedimientos automáticos de selección de variables, incluyendo la regresión paso a paso, es que operan sólo con base en la bondad del ajuste, y pueden no considerar las relaciones entre variables independientes, que son importantes.

Apéndice D

Análisis de Supervivencia con R

En el lenguaje R, el Análisis de Supervivencia se puede hacer a través de un conjunto de paquetes especializados. El principal paquete para realizar Análisis de Supervivencia es el paquete *survival* [5], otro paquete que también es útil es *fitdistrplus*. En este apéndice se describen (sin profundizar) algunas funciones de estos paquetes.

D.1. Paquete *survival*

El paquete *survival* permite llevar a cabo Análisis de Supervivencia para datos que presentan diversos mecanismos de censura. Para ejecutar cualquiera de las funciones de este paquete es necesario invocar la librería mediante la instrucción: *library(survival)* [5].

Función *Surv*

La función *Surv* permite crear un objeto de tipo *survival*, que usualmente se utiliza como una variable de respuesta en la fórmula de un modelo. La estructura para datos que presentan censura por la derecha es: *Surv(time, event)*, donde *time* es un vector que contiene los tiempos de seguimiento y *event* es otro vector con los indicadores de estado, normalmente 0 = censurado por la derecha y 1 = tiempo de falla.

Una estructura útil para estudiar datos que presentan otros tipos de censura es: *Surv(time1, time2, event, type="interval2")*. En este caso, se piensa que cada observación es un intervalo de tiempo de la siguiente manera: $(-\infty, t)$

para censura por la izquierda, (t, ∞) para censura por la derecha, (t, t) para tiempo de falla exacto y (t_1, t_2) para censura por intervalo. El argumento *time1* es un vector que contiene los extremos izquierdos de los intervalos observados, mientras que *time2* contiene los extremos derechos. Los valores infinitos se pueden representar Inf o por NA. Los indicadores de estado que contiene el vector *event* son: 0 = censura por la derecha, 1 = tiempo de falla, 2 = censura por la izquierda y 3 = censura por intervalo.

Función *survfit*

La función *survfit* permite obtener la estimación de la función de supervivencia utilizando el método de Kaplan y Meier (opción por defecto) o de Fleming y Harrington [5]. La estructura de la función *survfit* para obtener el estimador de Kaplan y Meier es: *survfit(Surv(...) $\sim$ 1)*, donde *Surv()* contiene sus respectivos argumentos. Si se requiere la estimación de la función de supervivencia para diferentes tratamientos se utiliza: *survfit(Surv(...) $\sim$ T)* donde *T* es un vector que indica a que tratamiento pertenece cada observación. La función de supervivencia estimada se obtiene mediante: *summary(survfit(...))*.

La gráfica de la función de supervivencia estimada con banda de confianza se obtiene con *plot(survfit(...), conf.int = T)* y sin banda de confianza con *plot(survfit(...), conf.int = F)*. Además, la gráfica de la función de riesgo acumulado estimada con banda de confianza se obtiene con *plot(survfit(...), fun = "cumhaz", conf.int = T)* y sin banda de confianza con *plot(survfit(...), fun = "cumhaz", conf.int = F)*.

Función *coxph*

La función *coxph* permite ajustar modelos de regresión de riesgo proporcional. La estructura de la función *coxph* para datos que presentan censura por la derecha es: *coxph(Surv(time, event)* $\sim x_1 + \ldots + x_k$ *)*, donde $x_1, \ldots, x_k$ son vectores que contienen los valores de las covariables (constantes) para cada observación.

Las instrucciones *print(coxph(...))* y *coxph(...)* permiten obtener los contrastes para verificar si el modelo de Cox ajustado es adecuado y con la instrucción *summary(coxph(...))* se obtiene un poco más de detalles de los contrastes. La función de supervivencia ajustada por el modelo de Cox se obtiene con *summary(survfit(coxph(...)))*, además, su gráfica se obtiene con *plot(survfit(coxph(...)))* [5].

D.2. Paquete *fitdistrplus*

El paquete *fitdistrplus* tiene como objetivo ayudar a ajustar distribuciones paramétricas univariadas a datos censurados o no censurados [10]. La librería se carga mediante la instrucción: *library(fitdistrplus)*.

Función *fitdist*

La función *fitdist* permite ajustas distribuciones univariadas a datos no censurados mediante la metodología de máxima verosimilitud [10]. La estructura de esta función es: *fitdist(time,distr)*, donde *time* es un vector que contiene los tiempos de falla y *distr* es el nombre de la distribución que se esta ajustando, por ejemplo “exp”, “weibull”, “lnorm” y “gamma”. Las estimaciones se pueden obtener mediante las siguientes instrucciones:
ajuste <- fitdist(...)
ajuste$estimate.

Función *fitdistcens*

Ajusta una distribución univariada a datos censurados mediante la metodología de máxima verosimilitud [10]. El ajuste se puede obtener mediante las siguientes instrucciones:
left <- time1
right <- time2
tiempo <- cbind.data.frame(left, right)
fitdistcens(tiempo,distr),
donde *time1* y *time2* son vectores que contiene los extremos de los intervalos observados como en la función *Surv* y *distr* es el nombre de la distribución que se esta ajustando como en la función *fitdist*. Las estimaciones se pueden obtener mediante las siguientes instrucciones:
ajuste <- fitdistcens(...)
ajuste$estimate.

Apéndice E

Salidas del *software* R

En este apéndice se presentan algunas de las salidas del *software* R obtenidas en el análisis de la deserción en las licenciaturas de la FCFM-BUAP (Capítulo 4).

Primero se presentan las matrices de correlación de las covariables de escala mínima de intervalo y después los resúmenes de los diferentes modelos de riesgo proporcional semiparamétricos:

- Con todas las covariables de escala mínima de intervalo.
- Obtenido mediante el procedimiento de regresión paso a paso con todas las covariables de escala mínima de intervalo.
- Obtenido mediante el procedimiento de regresión paso a paso con todas las covariables.

Por último, se muestran las gráficas de la función de supervivencia y de la función de riesgo acumulado ajustadas por los modelos de riesgo proporcional semiparamétrico mediante el procedimiento de regresión paso a paso para la LA, la LM y la LMA.

	Puntaje	Autoestima	HabEstudio	Lawson	THLB	Activo	Reflexivo	Teórico	Pragmático
Puntaje	1	0	0.14	0.35	0.17	-0.11	0.09	0.22	0.07
Autoestima		1	0.13	0.09	0.03	0.19	-0.12	-0.07	0.04
HabEstudio			1	-0.06	0	-0.01	0.16	0.4	0.21
Lawson				1	0.28	0.04	0.07	0.06	0.13
THLB					1	-0.07	-0.05	-0.1	-0.06
Activo						1	-0.36	-0.29	0.22
Reflexivo							1	0.48	0.12
Teórico								1	0.3
Pragmático									1

Tabla E.1: Correlación de las covariables de escala mínima de intervalo en la licenciatura en Actuaría, n=275.

	Puntaje	Autoestima	HabEstudio	Lawson	THLB	Activo	Reflexivo	Teórico	Pragmático
Puntaje	1	-0.06	0.03	0.58	0.41	-0.15	0.16	0.22	-0.16
Autoestima		1	0.14	0	0.04	0.02	0.01	-0.03	0.02
HabEstudio			1	-0.1	0.06	0.02	0.35	0.42	0.14
Lawson				1	0.31	-0.08	0.04	0.16	-0.08
THLB					1	-0.2	0.11	0.12	-0.04
Activo						1	-0.11	-0.14	0.29
Reflexivo							1	0.48	0.18
Teórico								1	0.41
Pragmático									1

Tabla E.2: Correlación de las covariables de escala mínima de intervalo en la licenciatura en Matemáticas, n=212.

	Puntaje	Autoestima	HabEstudio	Lawson	THLB	Activo	Reflexivo	Teórico	Pragmático
Puntaje	1	0.09	0.12	0.42	0.38	-0.19	0.02	0	0
Autoestima		1	0.12	0.1	0.12	-0.12	0.04	0.06	-0.04
HabEstudio			1	-0.03	0.02	-0.04	0.12	0.25	0.07
Lawson				1	0.2	-0.13	0.07	0.05	0.04
THLB					1	-0.23	0.02	-0.03	-0.1
Activo						1	-0.07	-0.12	0.25
Reflexivo							1	0.51	0.35
Teórico								1	0.43
Pragmático									1

Tabla E.3: Correlación de las covariables de escala mínima de intervalo en la licenciatura en Matemáticas Aplicadas, n=190.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
Puntaje	-0.006	0.994	0.003	-2.225	0.026	*
Autoestima	-0.035	0.965	0.019	-1.837	0.066	.
HabEstudio	-0.182	0.834	0.189	-0.966	0.334	
Lawson	-0.095	0.909	0.071	-1.352	0.176	
THLB	0.233	1.263	0.142	1.640	0.101	
Activo	0.041	1.042	0.042	0.992	0.321	
Reflexivo	-0.050	0.951	0.050	-1.012	0.312	
Teórico	-0.053	0.949	0.052	-1.013	0.311	
Pragmático	-0.028	0.972	0.050	-0.562	0.574	

Tabla E.4: Resumen del modelo de riesgo proporcional semiparamétrico con todas las covariables de escala mínima de intervalo para la Licenciatura en Actuaría, n= 275 y número de eventos= 67.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
Puntaje	-0.006	0.994	0.002	-3.358	0.001	***
Autoestima	-0.007	0.993	0.017	-0.393	0.694	
HabEstudio	-0.043	0.958	0.132	-0.322	0.747	
Lawson	0.038	1.039	0.059	0.649	0.516	
THLB	-0.021	0.979	0.096	-0.221	0.825	
Activo	0.010	1.010	0.036	0.283	0.777	
Reflexivo	-0.046	0.955	0.041	-1.123	0.262	
Teórico	-0.014	0.986	0.046	-0.316	0.752	
Pragmático	0.038	1.039	0.045	0.841	0.400	

Tabla E.5: Resumen del modelo de riesgo proporcional semiparamétrico con todas las covariables de escala mínima de intervalo para la Licenciatura en Matemáticas, n= 212 y número de eventos= 98.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
Puntaje	-0.007	0.993	0.002	-4.029	5.6e-05	***
Autoestima	-0.021	0.980	0.014	-1.428	0.153	
HabEstudio	-0.042	0.959	0.132	-0.321	0.748	
Lawson	0.108	1.114	0.056	1.930	0.054	.
THLB	0.000	1.000	0.090	0.001	0.999	
Activo	0.001	1.001	0.035	0.025	0.980	
Reflexivo	-0.021	0.980	0.045	-0.460	0.646	
Teórico	-0.010	0.990	0.049	-0.203	0.840	
Pragmático	-0.018	0.982	0.042	-0.431	0.667	

Tabla E.6: Resumen del modelo de riesgo proporcional semiparamétrico con todas las covariables de escala mínima de intervalo para la Licenciatura en Matemáticas Aplicadas, n= 190 y número de eventos= 99.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
Teórico	-0.118	0.888	0.038	-3.107	0.002	**
Puntaje	-0.006	0.994	0.002	-2.438	0.015	*
Autoestima	-0.034	0.967	0.019	-1.820	0.069	.

Signif. codes: 0 '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1

	exp(coef)	exp(-coef)	lower .95	upper .95
Teórico	0.888	1.126	0.824	0.957
Puntaje	0.994	1.006	0.990	0.999
Autoestima	0.967	1.035	0.932	1.003

Concordance= 0.657 (se = 0.042)
Rsquare= 0.072 (max possible= 0.92)

Likelihood ratio test	=	20.52	on 3 df,	p=0.0001325
Wald test	=	20.32	on 3 df,	p=0.0001454
Score (logrank) test	=	20.83	on 3 df,	p=0.0001144

Tabla E.7: Resumen del modelo de riesgo proporcional semiparamétrico mediante el procedimiento de regresión paso a paso con todas las covariables de escala mínima de intervalo para la Licenciatura en Actuaría, n= 275 y número de eventos= 67.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
Puntaje	-0.006	0.994	0.001	-4.45	8.6e-06	***

Signif. codes: 0 '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1

	exp(coef)	exp(-coef)	lower .95	upper .95
Puntaje	0.994	1.006	0.992	0.997

Concordance= 0.626 (se = 0.037)
Rsquare= 0.097 (max possible= 0.989)

Likelihood ratio test	=	21.66	on 1 df,	p=3.255e-06
Wald test	=	19.8	on 1 df,	p=8.605e-06
Score (logrank) test	=	20.34	on 1 df,	p=6.491e-06

Tabla E.8: Resumen del modelo de riesgo proporcional semiparamétrico mediante el procedimiento de regresión paso a paso con todas las covariables de escala mínima de intervalo para la Licenciatura en Matemáticas, n= 212 y número de eventos= 98.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
Puntaje	-0.008	0.993	0.002	-4.353	1.34e-05	***
Lawson	0.097	1.102	0.052	1.857	0.063	.

Signif. codes: 0 '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1

	exp(coef)	exp(-coef)	lower .95	upper .95
Puntaje	0.993	1.008	0.989	0.996
Lawson	1.102	0.908	0.995	1.221

Concordance= 0.647 (se = 0.039)
Rsquare= 0.104 (max possible= 0.993)
Likelihood ratio test = 20.89 on 2 df, p=2.906e-05
Wald test = 19.48 on 2 df, p=5.902e-05
Score (logrank) test = 19.8 on 2 df, p=5.026e-05

Tabla E.9: Resumen del modelo de riesgo proporcional semiparamétrico mediante el procedimiento de regresión paso a paso con todas las covariables de escala mínima de intervalo para la Licenciatura en Matemáticas Aplicadas, n= 190 y número de eventos= 99.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
MatRep	0.535	1.707	0.153	3.493	0.001	***
Teórico	-0.093	0.912	0.038	-2.451	0.014	*
Lawson	-0.164	0.849	0.064	-2.567	0.010	*
TipoBach	-1.071	0.343	0.522	-2.050	0.040	*

Signif. codes: 0 '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1

	exp(coef)	exp(-coef)	lower .95	upper .95
MatRep	1.707	0.586	1.265	2.304
Teórico	0.912	1.097	0.846	0.982
Lawson	0.849	1.178	0.749	0.962
TipoBach	0.343	2.917	0.123	0.954

Concordance= 0.701 (se = 0.042)
Rsquare= 0.11 (max possible= 0.92)
Likelihood ratio test = 31.59 on 4 df, p=2.319e-06
Wald test = 31.44 on 4 df, p=2.485e-06
Score (logrank) test = 34.46 on 4 df, p=5.991e-07

Tabla E.10: Resumen del modelo de riesgo proporcional semiparamétrico mediante el procedimiento de regresión paso a paso con todas las covariables para la Licenciatura en Actuaría, n=270 y número de eventos=66.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
Puntaje	-0.006	0.994	0.001	-4.136	3.53e-05	***
Trabajo	0.846	2.331	0.224	3.786	0.000	***
OpCarrera	0.717	2.048	0.273	2.627	0.009	**
TipoBach	-0.589	0.555	0.336	-1.750	0.080	.

Signif. codes: 0 '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1

	exp(coef)	exp(-coef)	lower .95	upper .95
Puntaje	0.994	1.006	0.992	0.997
Trabajo	2.331	0.429	1.504	3.612
OpCarrera	2.048	0.488	1.200	3.495
TipoBach	0.555	1.802	0.287	1.073

Concordance= 0.678 (se = 0.037)
Rsquare= 0.186 (max possible= 0.989)

Likelihood ratio test	=	43.09	on 4 df,	p=9.925e-09
Wald test	=	41.88	on 4 df,	p=1.765e-08
Score (logrank) test	=	44.57	on 4 df,	p=4.887e-09

Tabla E.11: Resumen del modelo de riesgo proporcional semiparamétrico mediante el procedimiento de regresión paso a paso con todas las covariables para la Licenciatura en Matemáticas, n= 209 y número de eventos= 97.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
Puntaje	-0.007	0.993	0.002	-4.205	2.61e-05	***
RecSem	-0.560	0.571	0.218	-2.568	0.010	*
FinBach	0.603	1.827	0.319	1.890	0.059	.
Lawson	0.095	1.100	0.052	1.836	0.066	.

Signif. codes: 0 '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1

	exp(coef)	exp(-coef)	lower .95	upper .95
Puntaje	0.993	1.007	0.989	0.996
RecSem	0.571	1.750	0.373	0.876
FinBach	1.827	0.547	0.978	3.414
Lawson	1.100	0.909	0.994	1.218

Concordance= 0.669 (se = 0.039)
Rsquare= 0.14 (max possible= 0.992)

Likelihood ratio test	=	27.99	on 4 df,	p=1.254e-05
Wald test	=	26.89	on 4 df,	p=2.094e-05
Score (logrank) test	=	27	on 4 df,	p=1.985e-05

Tabla E.12: Resumen del modelo de riesgo proporcional semiparamétrico mediante el procedimiento de regresión paso a paso con todas las covariables para la Licenciatura en Matemáticas Aplicadas, n= 186 y número de eventos= 95.

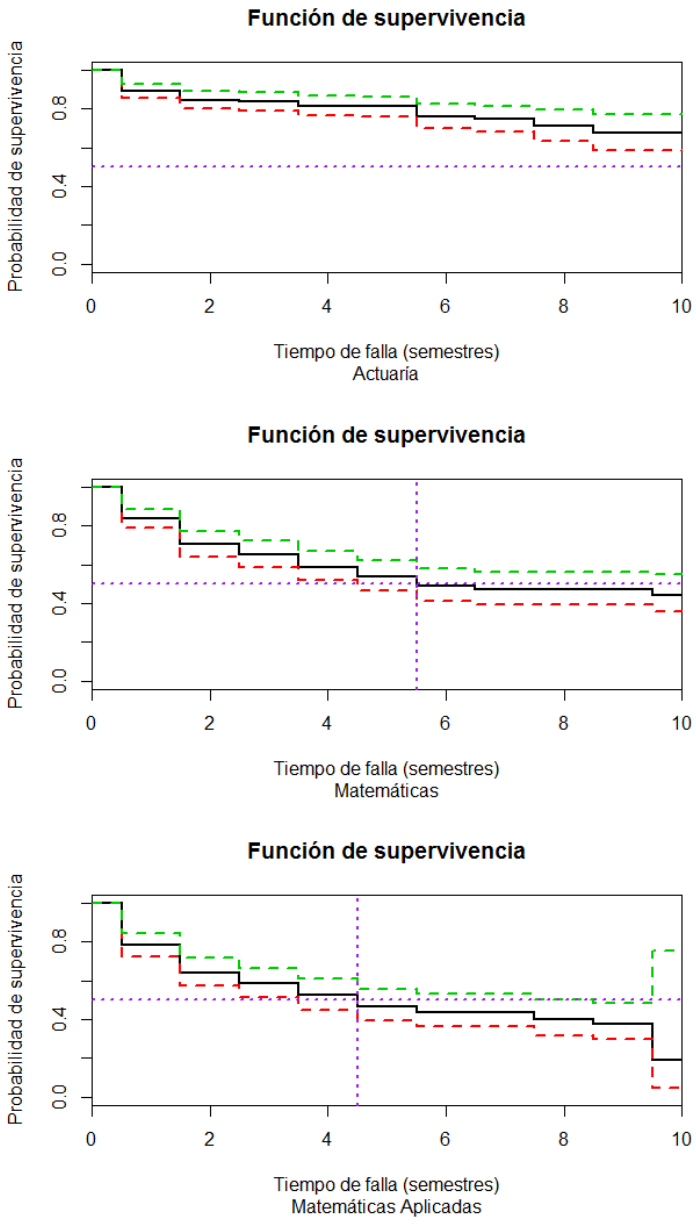

Figura E.1: Función de supervivencia ajustada por el modelo de riesgo proporcional semiparamétrico mediante el procedimiento de regresión paso a paso con todas las covariables de escala mínima de intervalo (con banda de confianza de 95 % y estimación de la mediana) para la LA, la LM y la LMA.

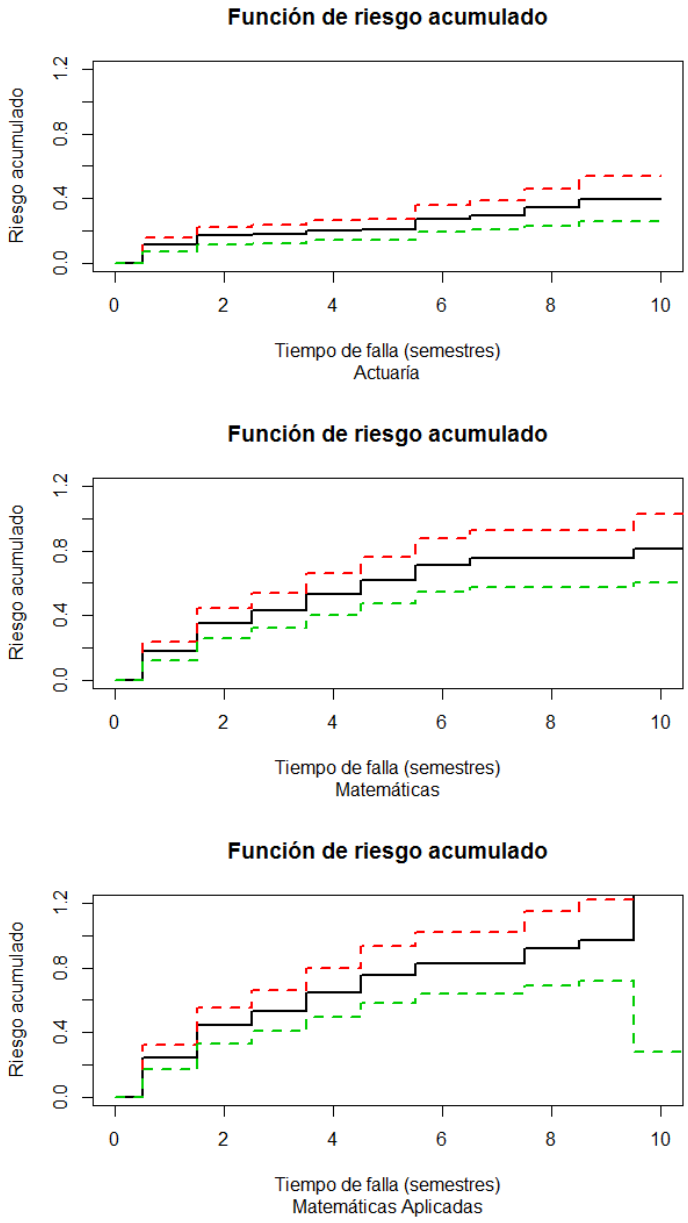

Figura E.2: Función de riesgo acumulado ajustada por el modelo de riesgo proporcional semi-paramétrico mediante el procedimiento de regresión paso a paso con todas las covariables de escala mínima de intervalo (con banda de confianza de 95 %) para la LA, la LM y la LMA.

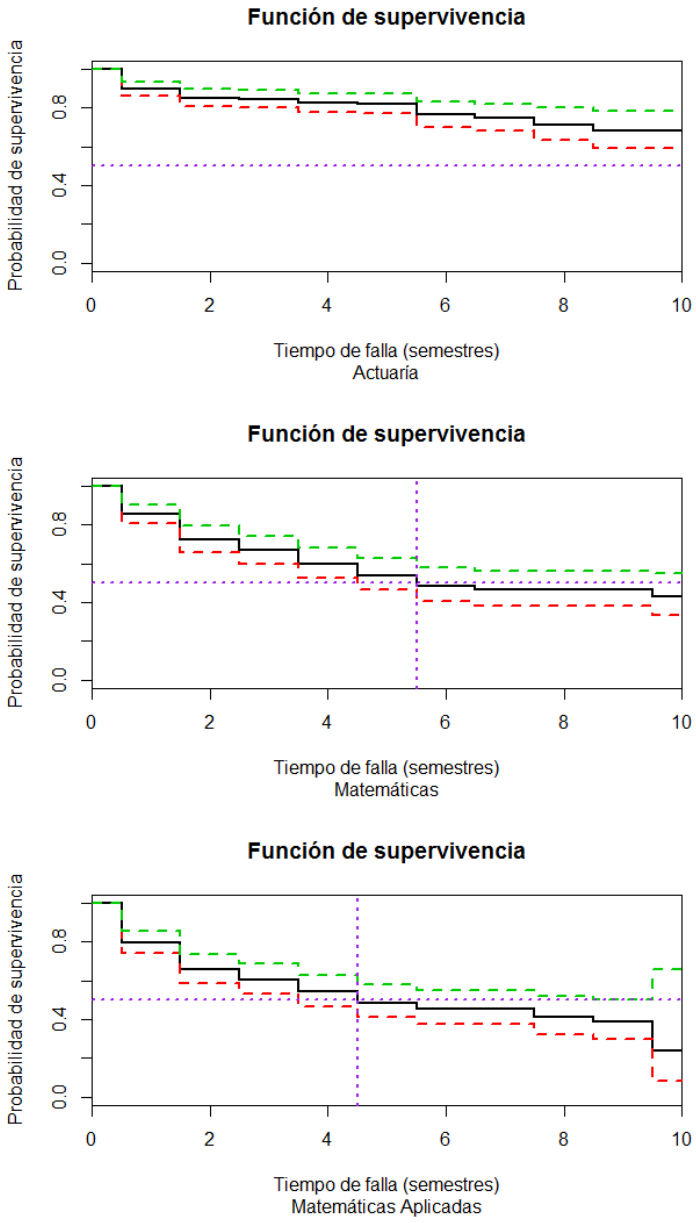

Figura E.3: Función de supervivencia ajustada por el modelo de riesgo proporcional semiparamétrico mediante el procedimiento de regresión paso a paso con todas las covariables (con banda de confianza de 95 % y estimación de la mediana) para la LA, la LM y la LMA.

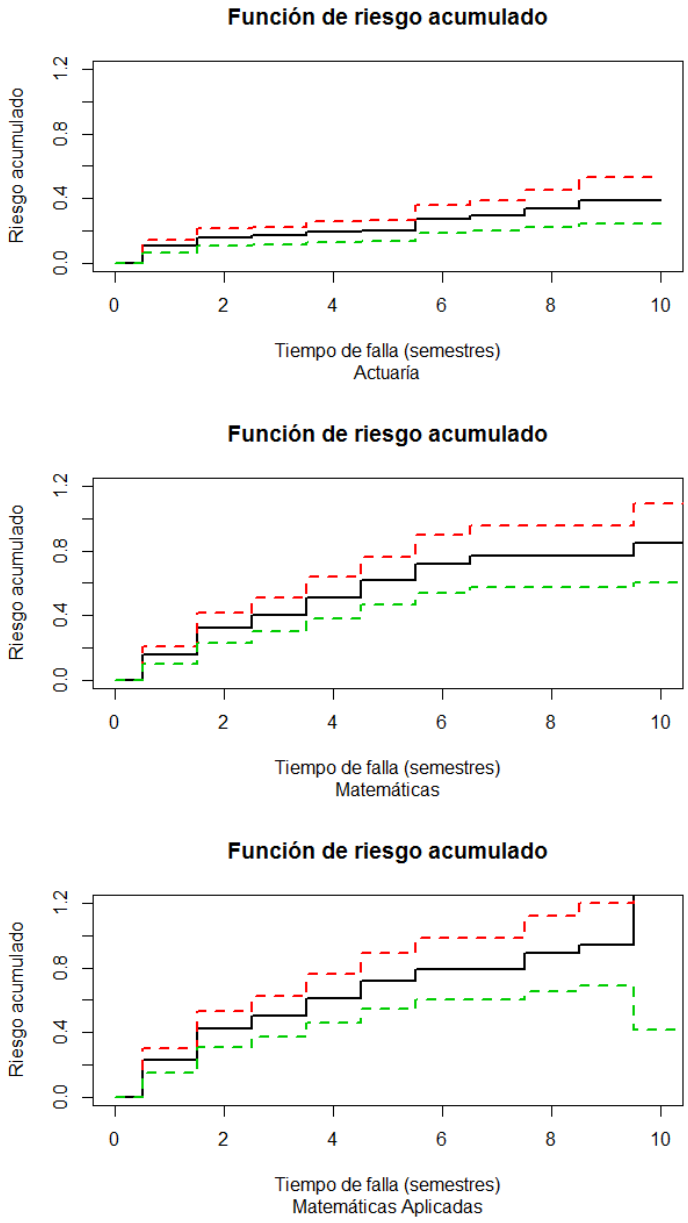

Figura E.4: Función de riesgo acumulado ajustada por el modelo de riesgo proporcional semiparamétrico mediante el procedimiento de regresión paso a paso con todas las covariables (con banda de confianza de 95 %) para la LA, la LM y la LMA.

Referencias

[1] Alonso, C.; Gallego D. J.; Honey, P. 1994. *Cuestionario de Honey-Alonso de Estilos de Aprendizaje*. Madrid: Instituto de Ciencias de la Educación (ICE).

[2] Alonso, C. M.; Gallego, D. J.; Honey, P. 1995. *Los estilos de aprendizaje. Procedimientos de diagnóstico y mejora*. 6^a ed., Bilbao, Ediciones Mensajero.

[3] Arenas Martinez, G. Y. *Una aplicación de regresión lineal en el aprovechamiento de los alumnos de nuevo ingreso en el área de matemáticas de la FCFM*. Tesis de licenciatura, Facultad de Ciencias Físico Matemáticas, Benemérita Universidad Autónoma de Puebla, 2011.

[4] Ates, S.; Cataloglu, E. 2007. *The effects of students'reasoning abilities on conceptual understandings and problem-solving skills in introductory mechanics*. European Journal of Physics, vol. 28, núm. 6, págs. 1161-1171.

[5] Borges Peña, R. E. 2005. *Análisis de Sobrevivencia utilizando el lenguaje R*. XV Simposio de Estadística, Universidad Nacional de Colombia, Paipa, Boyacá, Colombia.

[6] Brinkmann Sch., H.; Segure M., T.; Solar R., Ma. I. 1989. *Adaptación, estandarización y elaboración de Normas para el Inventario de Autoestima de Coopersmith*. Rev. Chilena de Psicología, Vol 10, num. 1, págs. 63-71, ISSN 0716-3630.

[7] CEDE. 2007. *Investigación sobre Deserción en las Instituciones de Educación Superior en Colombia*. Universidad de los Andes, Bogotá, Colombia.

[8] Coopersmith, S. 1967. *The antecedents of self-esteem*. San Francisco: W.H. Freeman.

[9] Cox, D. R.; Oakes, D. 1984. *Analysis of Survival Data*. 1^a ed., Gran Bretaña, Chapman & Hall.

[10] Delignette Muller, M. L.; Dutang, C.; Pouillot, R.; Denis, J. B.; Siberchicot, A. *fitdistrplus: Help to Fit of a Parametric Distribution to Non-Censored or Censored Data*. R package versión 1.0-9. https://cran.r-project.org/web/packages/fitdistrplus/index.html. Consultado en Agosto de 2017.

[11] Díaz Jiménez, H. J. 2012. *Estudio comparativo entre el aprendizaje colaborativo y el tradicional en la enseñanza de los conceptos de calor y temperatura a nivel medio superior*. Tesis de maestría, Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada, Instituto Politécnico Nacional.

[12] Díaz Ramírez, J. *Análisis Descriptivo de los Egresados y Titulados de las Licenciaturas de Matemáticas y Matemáticas Aplicadas de las Generaciones 2000 a 2004*. Tesis de licenciatura, Facultad de Ciencias Físico Matemáticas, Benemérita Universidad Autónoma de Puebla, 2013.

[13] Freiberg Hoffman, A.; Fernández Liporace, M. 2013. *Cuestionario Honey-Alonso de estilos de aprendizaje: Análisis de sus propiedades Psicométricas en Estudiantes Universitarios*. Summa Psicológica UST, vol. 10, núm. 1, págs. 103-117, ISSN 0718-0446.

[14] Gill, R. D. 2001. *Product Integration*. Mathematical Institute, University of Utrecht, Netherlands. Eurandom, Eindhoven, Netherlands.

[15] Hernández Guerra, S. *Uso del Modelo de Regresión Logística para Estudiar la aprobación de la materia de Matemáticas Básicas de la FCFM en las Generaciones 2010 y 2011*. Tesis de licenciatura, Facultad de Ciencias Físico Matemáticas, Benemérita Universidad Autónoma de Puebla, 2013.

[16] "Historia Universitaria". BUAP. http://www. buap. mx/. Consultado: 14 de abril de 2017.

[17] Honey, P.; Mumford, A. 1986. *The Manual of Learning Styles*. Maidenhead, Berkshire: P. Honey, Ardingly House.

[18] Juárez Hernández, B.; Izquierdo Valladares, O. M. 2010. Gama generalizada, una súper familia de distribuciones "en análisis de supervivencia". *Aportaciones y aplicaciones de la probabilidad y la estadística*. FCFM - Benemérita Universidad Autónoma de Puebla, págs. 88-108.

[19] Kalbfleisch, J. D.; Prentice, R. L. 2002. *The Statistical Analysis of Failure Time Data*. 2^a ed., Nueva Jersey, Wiley-Interscience.

[20] Keefe, J. 1988. *Aprendiendo Perfiles de Aprendizaje*. Asociación Nacional de Escuelas Secundarias.

[21] Khodabin, M.; Ahmadabadi, A. 2010. *Some properties of generalized gamma distribution*. Mathematical Sciences, vol. 4, núm. 1, págs. 9-28.

[22] Klein, J. P.; Moeschberger, M. L. 2003. *Survival Analysis. Techniques for Censored and Truncated Data*. 2^a ed., Nueva York, Springer.

[23] Kolb, D. 1984. *Experiential learning: experience as the source of learning and development*. Nueva Jersey: Prentice Hall, Inc., Englewood Cliffs.

[24] Lara Cantú, Ma. A.; Verduzco, Ma. A.; Acevedo, M.; Cortés, J. 1993. *Validez y confiabilidad del inventario de autoestima de Cooper Smith para adultos, en población mexicana*. Revista Latinoamericana de Psicología (en línea), vol. 25, núm. 2, págs. 247-255. Consultado: Agosto de 2017. Disponible en: http://www.redalyc.org/articulo.oa?id=80525207>. ISSN 0120-0534.

[25] Lawless, J. F. 2003. *Statistical Models and Methods for Lifetime Data*. 2^a ed., Nueva Jersey, Wiley-Interscience.

[26] Maldonado García, A. *Identificación de factores que intervienen en la reprobación del curso de Matemáticas Básicas de la FCFM de la BUAP*. Tesis de licenciatura, Facultad de Ciencias Físico Matemáticas, Benemérita Universidad Autónoma de Puebla, 2012.

[27] Mare, R. D. 1980. *Social Background and School Continuation Decision*. Journal of the American Statistical Association, vol. 75, núm. 370, págs. 295-305.

[28] Montes Gutiérrez, I. C.; Almonacid Hurtado, P. Ma.; Gómez Cardona, S.; Zuluaga Díaz, Fco. I.; Tamayo Zea, E. 2010. *Análisis de la deserción estudiantil en los programa de pregrado de la Universidad EAFIT*. Grupo de investigación estudios en economía y empresa, Departamento de Economía, Escuela de Administración, Universidad EAFIT. Medellín. ISSN 1692-0694.

[29] Nieto Méndez, A. L. *Estimación de la probabilidad de egreso de estudiantes de licenciatura en ciencias de la BUAP usando Regresión Logística*.

Tesis de licenciatura, Facultad de Ciencias Físico Matemáticas, Benemérita Universidad Autónoma de Puebla, 2015.

[30] Pinzón Rosas, A. J. 2011. *Comparación del aprendizaje, la actitud y el razonamiento científico en la enseñanza de la química a distancia y en la enseñanza tradicional*. Revista de investigaciones UNAD, vol. 10, núm. 1, págs. 167-173. ISSN 0124 793X.

[31] "Requisitos para nuevo ingreso: Licenciatura y Técnico Superior Universitario. Modalidad Escolarizada". BUAP. http://www.admision.buap.mx/sites/default/files/licenciatura.pdf. Consultado en Agosto de 2017.

[32] Riquelme del Solar, G.; Constenla Núñez, J.; Cruces Escobar, O. 1996. *Test de habilidades Lectocomprensivas Básicas, "THLB", versión experimental*. Estudios Pedagógicos, núm 22, págs. 111-127.

[33] Rodríguez Lagunas, J.; Leyva Piña, M. A. 2007. *La deserción escolar universitaria. La experiencia de la UAM. Entre el déficit de la oferta educativa superior y las dificultades de la retención escolar*. El Cotidiano, vol. 22, núm. 142, págs. 98-111.

[34] Rondon, C. 1991. *Internalidad y Hábitos de Estudio*. Tesis de Maestría. Universidad Pedagógica Experimental Libertador, Instituto Pedagógico de Barquisimeto, Venezuela.

[35] Silva Laya, M. 2011. *El primer año universitario. Un tramo crítico para el éxito académico*. Perfiles Educativos, vol. XXXIII, número especial, págs. 102-114.

[36] Smith, J. P.; Naylor, R. A. 2001. *Dropping Out of University: A Statistical Analysis of the Probability of Withdrawal for UK University Students*. Journal of the Royal Statistical Society, Series A (Statistics in Society), vol. 164, núm. 2, págs. 389-405.

[37] Solano, E. 2006. *Causas e Indicadores de la Deserción en el Programa de Economía de la Universidad del Atlántico Aplicando Modelos de Duración y Microeconómico*. Universidad del Atlántico, Barranquilla.

[38] Stratton, L. S.; O'Toole, D. M.; Wetzel, J. N. 2005. *A Multinomial Logit Model of College Stopout and Dropout Behavior*. IZA - Institute for the Study of Labor. Bonn: IZA.

[39] Symanzik, J. *Mathematical Statistics I*. Utah State University, Department of Mathematics and Statistics. http://www.math.usu.edu/∼symanzik/teaching/1999_stat6710/lect_main.pdf. Consultado en Agosto de 2017.

[40] Tinto, V. 1987. *El abandono de los estudios superiores: una nueva perspectiva de las causas del abandono y su tratamiento*. México, UNAM.

[41] Tinto, V. 1987. *Leaving College: Rethinking the Causes and Cures of Student Attrition*. 2^a ed., Chicago, University of Chicago Press.

[42] Universidad Nacional. 2007. *Cuestión de Supervivencia*. Bogotá, Universidad Nacional de Colombia.

[43] Vásquez, J.; Castaño, E.; Gallón, S.; Gómez, K. 2003. *Determinantes de la Deserción Estudiantil en la Universidad de Antioquia*. Universidad de Antioquia, Facultad de Ciencias Económicas, Medellín.

[44] Zwolak, J. P.; Dou, R.; Williams, E. A.; Brewe, E. 2017. *Student' network integration as a predictor of persistence in introductory physics courses*. Physical Review Physics Education Research, vol. 13, núm. 1, 010113.