Análisis estadístico en datos sobrevida

Luis Guillermo Díaz

Facultad de Ciencias
Sede Bogotá

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Análisis estadístico en datos de sobrevida

Análisis estadístico en datos de sobrevida

Luis Guillermo Díaz Monroy

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Bogotá D. C., Colombia, 2025

Primera edición, 2025
ISBN 978-628-503-044-4 (digital)

Edición
Alejandro Cano Moreno
Coordinación de Publicaciones, Facultad de Ciencias, Sede Bogotá
coorpub_fcbog@unal.edu.co

Corrección de estilo
Johny Andrés Martínez Cano

Diseño de carátula
Damián Crofort

Maqueta LaTeX
Camilo Cubides

Editado en Bogotá, D. C., Colombia

Catalogación en la publicación Universidad Nacional de Colombia

Díaz Monroy, Luis Guillermo, 1958-
Análisis estadístico en datos de sobrevida / Luis Guillermo Díaz Monroy. -- Primera edición. -- Bogotá : Universidad Nacional de Colombia. Facultad de Ciencias : Coordinación de publicaciones Facultad de Ciencias, 2025.
1 recurso en línea (439 páginas) : ilustraciones en blanco y negro, diagramas. -- (Colección Textos. Estadística)

Incluye referencias bibliográficas
ISBN 978-628-503-044-4 (digital)

1. Análisis de supervivencia (Biometría) -- Modelos matemáticos 2. Análisis de datos -- Métodos estadísticos 3. Supervivencia -- Modelos matemáticos 4. Esperanza de vida -- Modelos matemáticos 5. Probabilidades -- Estudio y enseñanza superior 6. Distribución (Teoría de probabilidades) 7. Estadística matemática -- Metodología -- Obras educativas 8. Procesos estocásticos 9. Estimación estadística 10. Datos aleatorios (Estadística) 11. Análisis de datos de tiempo de falla 12. Modelos de riesgos proporcionales 13. Riesgos en la competencia 14. Contrastación de hipótesis (Estadística) 15. Análisis multivariante 16. Análisis de regresión 17. Teoría bayesiana de decisiones estadísticas 18. Estadística matemática -- Proceso de datos 19. Paquetes de R I. Título II. Serie

CDD-23 519.546 / 2025

Contenido

Lista de figuras VII
Lista de tablas XI
Prefacio XV

Capítulo uno
Contexto y elementos generales del análisis de sobrevivencia 1
1.1. Introducción 3
1.2. Estructura de los datos de sobrevida 8
1.2.1. Variable aleatoria *tiempo hasta el evento* 13
1.2.2. Censura y truncamiento 14
1.2.3. Censura a la derecha Tipo I 16
1.2.4. Censura a la derecha Tipo II 17
1.2.5. Censura a la derecha bajo riesgos en competencia 17
1.2.6. Censura a la izquierda 18
1.2.7. Censura por intervalo 19
1.3. Censura aleatoria (Tipo III) 20
1.3.1. Truncamiento 21
1.4. Función de sobrevida 23
1.5. Función de riesgo 28

Capítulo dos
Métodos no paramétricos para estimar la función de sobrevida 39
2.1. Introducción 41
2.2. Caso sin censura 41
2.3. Estimador de Kaplan-Meier 42
2.4. Bandas de confianza para la función de sobrevida 47
2.5. Otros estimadores no paramétricos 50
2.5.1. Estimador de Nelson-Aalen 50
2.6. Estimador de $h(t)$ mediante el suavizador de núcleo 53
2.6.1. Estimador de tabla de vida actuarial 56
2.7. Comparación de funciones de sobrevida 59

2.7.1. Estadística log-rank 61
2.7.2. Otras estadísticas (Wilcoxon y Tarone-Ware) 67

Capítulo tres
Modelos probabilísticos en análisis de sobrevida **69**
3.1. Introducción 71
3.2. Distribución exponencial 71
3.3. Distribución exponencial generalizada 72
3.4. Distribución de Weibull 74
3.5. Distribución de log-normal 78
3.6. Distribución log-logística 80
3.7. Distribución gama 84
3.8. Distribución gama generalizada 88
3.9. Distribución de valor extremo 89
3.10. Distribución de Birnbaum-Saunders 90
3.11. Otras distribuciones de sobrevida 93

Capítulo cuatro
Estimación de parámetros en modelos de sobrevida **97**
4.1. Introducción 99
4.2. Construcción de la función de verosimilitud 100
4.2.1. Verosimilitud para datos con censura Tipo I 104
4.2.2. Verosimilitud para datos truncados 107
4.2.3. Verosimilitud para datos con censura Tipo II 109
4.2.4. Verosimilitud para datos con censura aleatoria 110
4.3. Ajuste a los datos de un modelo probabilístico 118
4.3.1. Prueba de bondad de ajuste 143

Capítulo cinco
Modelos de regresión paramétricos en A. S. **147**
5.1. Introducción 149
5.2. Modelos de riesgos proporcionales 150
5.3. Modelos de tiempo acelerado 153
5.3.1. Modelo lineal para datos de tiempo hasta un evento 157

Capítulo seis
Modelo de riesgos proporcionales de Cox **173**
6.1. Introducción 175
6.2. Modelo de Cox 176

6.2.1. Variables del componente lineal 177
6.3. Estimación en el modelo de Cox 181
6.3.1. Tratamiento de empates 186
6.3.2. Interpretación de los coeficientes β's 188
6.3.3. Estimación de funciones relacionadas con $h_0(t)$ 190
6.4. Adecuación del modelo de Cox 196
6.4.1. Métodos gráficos y residuos 196
6.4.2. Comparación de modelos alternativos 199

Capítulo siete
Extensiones del modelo de Cox **225**
7.1. Introducción 227
7.2. Modelos con covariables tiempo dependientes 228
7.3. Modelos estratificados 239
7.4. Modelo de Cox con censura por intervalo 243
7.5. Modelos de fragilidad o vulnerabilidad 249

Capítulo ocho
Modelos de riesgo aditivos **261**
8.1. Introducción 263
8.2. Modelo de riesgo aditivo de Aalen no paramétrico 263
8.3. Estimación 265
8.4. Modelo de riesgo aditivo de Lin-Ying 271

Capítulo nueve
Análisis bayesiano en datos de sobrevida **275**
9.1. Introducción 277
9.2. Elementos del análisis estadístico bayesiano 278
9.3. Muestreo desde la distribución *a posteriori* 281
9.4. Modelos de regresión paramétricos 282
9.5. Modelos de regresión semiparamétricos 292
9.6. Una generalización del modelo de Cox 295

Capítulo diez
Modelos para datos de sobrevida discretos **301**
10.1. Introducción 303
10.2. Modelos de regresión discretos 305
10.2.1. Modelos de regresión paramétricos 306
10.3. Modelo de riesgos proporcionales discreto 309

Capítulo once
Tamaño de muestra en análisis de sobrevida **319**
11.1. Introducción 321
11.2. Potencia y tamaño de muestra 322
11.3. Determinación de la probabilidad del evento 325
11.4. Tamaño de muestra para comparar dos distribuciones 327
11.5. Tamaño de muestra en dos distribuciones 328
11.6. Probabilidad del evento desde una estimación no paramétrica de la función de sobrevida 329

Capítulo doce
Modelamiento conjunto de datos longitudinales y de sobrevida **333**
12.1. Introducción 335
12.2. Modelamiento de datos longitudinales 336
12.3. Modelamiento conjunto en datos longitudinales y de sobrevida 343
12.3.1. DeGruttola y Tu (1994) 345
12.3.2. Tsiatis, DeGruttola y Wulfsohn (1995) 347
12.3.3. Wulfsohn y Tsiatis (1997) 349
12.3.4. Henderson, Diggle y Dobson (2000) 351

Capítulo trece
Datos de sobrevida multivariados **355**
13.1. Introducción 357
13.1.1. Datos paralelos 364
13.1.2. Datos longitudinales 364
13.2. Distribuciones de sobrevida multivariadas 366
13.3. Distribución exponencial bivariada de Gumbel 368
13.4. Modelos multivariados de sobrevida 368
13.4.1. Distribución log-normal multivariada 368
13.4.2. Distribución exponencial bivariada en la versión de Freund (1961) 369
13.4.3. Distribución exponencial bivariada en la versión de Marshall-Olkin (1967) 370
13.5. Fragilidad 370
13.5.1. Algunos generadores de distribuciones de fragilidad 371
13.6. Estructura de cópulas 376
13.7. Modelos de fragilidad o vulnerabilidad 378
13.7.1. Modelo de vulnerabilidad compartida para datos recurrentes 380

13.7.2. Análisis de eventos recurrentes 386
13.8. Modelos multiestado 391

Apéndice A
Procedimientos básicos con R **395**
A.1. Lectura de datos externos 397
A.1.1. El directorio de trabajo 397
A.1.2. Lectura de datos desde un archivo de texto 398
A.1.3. Lectura de datos desde un archivo CSV 398
A.1.4. Lectura de datos desde un archivo de Excel 400
A.2. Selección y transformación de datos 400
A.2.1. Creación de nuevas variables 401
A.2.2. Selección de subconjuntos de un marco de datos 401
A.2.3. Cálculos por niveles de un factor 403
A.3. Cálculo de probabilidades y cuantiles 405
A.3.1. Distribución binomial 405
A.3.2. Distribución de Poisson 406
A.3.3. Distribuciones normal y ji-cuadrado 407

Referencias **413**

Lista de figuras

1.1. Tiempo al evento en estudio sobre cinco unidades. 10
1.2. Tiempo al evento en cinco unidades con origen común. . . . 11
1.3. Diagrama de Lexis para censura generalizada Tipo I. 12
1.4. Línea de tiempo. 14
1.5. Función de sobrevida o de confiabilidad. 24
1.6. Funciones de sobrevida continua y discreta. 25
1.7. Funciones de sobrevida. 26
1.8. Función de sobrevida discreta. 28
1.9. Funciones de riesgo. 29
1.10. Función de riesgo tipo "bañera". 31
1.11. Función de probabilidad, sobrevida y riesgo en exp(θ). 33
1.12. Cuantiles. 35
1.13. Media residual y esperanza de vida a la edad t_0. 36

2.1. Función de sobrevida discreta. 42
2.2. Construcción de intervalos en estimador Kaplan–Meier. . . . 43
2.3. Estimadores de KM y NA para la función de sobrevida. . . . 53
2.4. Función de riesgo de tiempo de sobrevida a cáncer de ovario, suavizada mediante núcleo. 57
2.5. Función de sobrevida en estudio de hepatitis. 58
2.6. Comparación de funciones de sobrevida. 60
2.7. Funciones de sobrevida: germinación fríjol. 65

3.1. Función de probabilidad exponencial, sobrevida y riesgo asociadas. 73
3.2. Función de probabilidad Weibull. 75
3.3. Función de sobrevida Weibull. 76
3.4. Función de riesgo Weibull. 77
3.5. Función de probabilidad log-normal. 78
3.6. Función de sobrevida log-normal. 79
3.7. Función de riesgo log-normal. 80
3.8. Función de densidad de probabilidad log-logística. 82

3.9. Función de sobrevida log-logística. 83
3.10. Función de riesgo log-logística. 84
3.11. Función de densidad de probabilidad gama. 86
3.12. Función de sobrevida gama. 87
3.13. Función de riesgo gama. 88
3.14. Funciones de densidad Birnbaum-Saunders. 92
3.15. Función de riesgo lineal-exponencial. 93
3.16. Función de riesgo Gompertz. 94

4.1. Información de probabilidad para la verosimilitud (evento E). 105
4.2. Funciones de probabilidad truncadas. 108
4.3. Escogencia de un modelo para los datos observados. 120
4.4. Función de sobrevida empírica. 122
4.5. Función de sobrevida empírica y cuantil muestral. 123
4.6. Gráfico de probabilidad: ajuste de distribución exponencial. 127
4.7. Gráfico de probabilidad: ajuste de distribución Weibull. . . . 130
4.8. Gráfico de probabilidad: ajuste de distribución log-normal. 133
4.9. Gráfico de riesgo: ajuste de la distribución exponencial. 136
4.10. Gráfico de riesgo: ajuste de la distribución Weibull. 137
4.11. Función de sobrevida log-normal. 138
4.12. Gráficas de la función de riesgo. 139
4.13. Gráfica de residuos de Cox-Snell para un modelo log-normal. 143

5.1. Ajuste de datos de la tabla 5.1. 164
5.2. Ajuste a la exponencial de datos de sobrevida por leucemia. 166
5.3. Gráficos de probabilidad para los datos de la tabla 5.1. 168
5.4. Curvas de sobrevida (leucemia) ajustadas al modelo exponencial. 171
5.5. Comparación de curvas de sobrevida (leucemia) ajustadas al modelo de regresión exponencial. 172

6.1. Gráfica de riesgos proporcionales y riesgos no proporcionales. 180
6.2. Tiempo al evento en cinco unidades. 185
6.3. Razón de tasas de falla $\exp\{\beta_1 + \beta_2 t\}$. 208
6.4. Residuos de Schoenfeld frente al tiempo por *estadio* y *edad*. 212
6.5. Residuos de Schoenfeld frente al tiempo por *estadio* y *edad*. 213
6.6. Funciones de sobrevida estimadas por el modelo de Cox para los datos de cáncer de laringe. 216

6.7. Riesgos acumulados estimados por el modelo de Cox para los datos de cáncer de laringe. 217

7.1. Log-bilirrubina (Lbrt) del paciente núm. 1. 235
7.2. Funciones de riesgo acumulado base estimadas por tratamiento de cirrosis. 239
7.3. Funciones de sobrevida estimadas por tratamiento de cirrosis. 240

8.1. Funciones de regresión acumulada del modelo 8.3.8. 270

9.1. Trazabilidad del Gibbs y densidad marginal posterior de β_0. 287
9.2. Trazabilidad del Gibbs y densidad marginal posterior de β_1. 287
9.3. Trazabilidad del Gibbs y densidad marginal posterior de β_0. 291
9.4. Trazabilidad del Gibbs y densidad marginal posterior de β_1. 292
9.5. Trazabilidad del Gibbs y densidad marginal posterior de β_2. 292
9.6. Trazabilidad del Gibbs y densidad marginal posterior de β_3. 293
9.7. Trazabilidad del Gibbs y densidad marginal posterior de β_4. 293
9.8. Función de riesgo base constante a trozos. 294
9.9. Trazabilidad del Gibbs y densidad marginal posterior de λ_1. 300
9.10. Trazabilidad del Gibbs y densidad marginal posterior de β_1. 300

13.1. Tipologías de modelos multiestado. 393

Lista de tablas

1.1. Tiempo hasta aprender a escribir 22
1.2. Esperanza de vida residual en t 37
1.3. Vida futura esperada a la edad t_0 37

2.1. Sobrevida en estudio de hepatitis 47
2.2. Sobrevida en estudio de hepatitis 47
2.3. Estimación de la sobrevida mediante Nelson-Aalen 52
2.4. Tiempos de sobrevida de mujeres con cáncer de ovario 56
2.5. Sobrevida en estudio de hepatitis 59
2.6. Núm. de eventos por grupo al momento $t_{(j)}$ 61
2.7. Extensión a g-grupos del evento al momento $t_{(j)}$ 63
2.8. Tiempo a la germinación en semillas de fríjol 64
2.9. Comparación de curvas de sobrevida: datos de fríjol 66
2.10. Curvas de sobrevida en datos de hepatitis 68

3.1. Algunos modelos probabilísticos para el A. S. 95

4.1. Componentes de probabilidad bajo censura Tipo I 107
4.2. Función de probabilidad bajo truncamiento 109
4.3. Remisión en estudio de leucemia aguda 126
4.4. Fundición de motores de refrigeración 129
4.5. Tiempo a la falla de baterías de teléfono móvil 132
4.6. Tiempo a la muerte de pacientes con leucemia aguda 135
4.7. Tiempo a la falla en pastillas de frenos 142

5.1. Datos de tiempo a la muerte por leucemia 163
5.2. Regresión del tiempo sobre CGB 165
5.3. Estimación de regresión exponencial de los datos de leucemia 170
5.4. Estimación de regresión exponencial de los datos de leucemia 170

6.1. Codificación de los niveles de un factor 178

6.2. Tiempos al evento 183
6.3. Tiempo de sobrevida en mujeres con cáncer de mama 192
6.4. Tiempo de sobrevida de pacientes que padecen melanoma . 194
6.5. Tiempo de sobrevida en pacientes con cáncer de próstata . . 204
6.6. Modelos ajustados para los datos de cáncer de próstata 205
6.7. Datos de pacientes con cáncer de laringe 219
6.8. Modelos de Cox para datos de pacientes con cáncer de laringe 220
6.9. Verificación de modelos con parámetros de interacción 220
6.10. Verificación de la proporcionalidad en el modelo de Cox con interacción *estadio* y *edad* 221
6.11. Verificación de la proporcionalidad en el modelo de Cox sin la interacción *estadio* y *edad* 221
6.12. Modelo de regresión de Cox ajustado para los datos de cáncer de laringe 222
6.13. Estimación de $\hat{S}_0(t)$ y $\hat{H}_0(t)$ para datos de cáncer de laringe 223

7.1. Tiempo de sobrevida de pacientes con cirrosis de hígado . . 233
7.2. Seguimiento de la bilirrubina en pacientes con cirrosis de hígado 234
7.3. Modelos sin variable tiempo-dependiente 235
7.4. Modelos con variable tiempo-dependiente 236
7.5. Estimación de $H_0(t)$ en estudio de cirrosis 238
7.6. Tiempos de sobrevida de pacientes con AML 242
7.7. Duración de la remisión en pacientes con leucemia aguda . . 255
7.8. Datos con variable tiempo-dependiente 258
7.9. Tiempos de sobrevida de pacientes con cáncer de mama . . . 259

9.1. Resumen posterior de β con el modelo exponencial 286
9.2. Resumen posterior de los β's con el modelo Weibull 291
9.3. Modelo exponencial a trozos en ratas con cáncer 298
9.4. Estimadores de máxima verosimilitud 299
9.5. Resúmenes e intervalos posteriores 299

10.1. Modelos de riesgo discretos 309

11.1. Núm. de eventos por grupo al momento $t_{(j)}$ 328

13.1. Tiempos al evento repetidos 358
13.3. Tiempos al evento repetidos 359
13.2. Tiempos en pares de respuesta: t_1: pre-droga y t_2: post-droga 360

13.4. Tiempos de uso de energía eléctrica para una lámpara 361
13.5. Conteo de episodios de epilepsia. Grupo placebo 362
13.6. Tiempo al tumor (en días) en ratas de laboratorio 363
13.7. Tiempo al tumor (en días) en ratas de laboratorio 365
13.8. Número de cópulas hasta primer y segundo embarazo 375
13.9. Número de reclamos en pólizas 385

A.1. Distribuciones de probabilidad con R 411

Prefacio

layout2.pdf

El análisis estadístico en datos de sobrevida es una colección de métodos estadísticos utilizados para direccionar las preguntas y las respuestas que tienen que ver con la posibilidad (el *si*) y el momento (el *cuándo*) en que ocurre un evento de interés sobre una unidad (objeto). El análisis de sobrevida, o más exactamente el análisis estadístico del tiempo hasta la ocurrencia de un evento, juega un papel importante en las ciencias de la salud, la epidemiología, la biología, la demografía, la economía, la ingeniería, la actuaria, las ciencias sociales y la pedagogía, entre otras áreas de investigación.

La inclusión de las palabras *sobrevida*, *sobrevivencia* e incluso *supervivencia* en el lenguaje estadístico tiene su origen en la búsqueda de respuestas a las preguntas acerca de la probabilidad de que un individuo, expuesto o considerado bajo algunas condiciones o factores específicos, esté vivo después de un determinado tiempo y dentro de un periodo de estudio o seguimiento. Históricamente, estas preguntas o inquietudes que están relacionadas con el tiempo que pasa hasta la muerte de un individuo (sobrevida) surgieron en la medicina y en áreas afines. De esta manera, el interés estadístico se dirige a modelar el tiempo de vida de una persona bajo algunas condiciones particulares. El evento de interés es la muerte, y se registra el tiempo inmediatamente anterior a su ocurrencia; a este tiempo se le denomina *sobrevivencia*, *sobrevida* o *supervivencia*, indistintamente. En estos vocablos el prefijo *sobre* o *super* significan *por encima* o *después de*, lo cual supone una medición del tiempo que transcurre después o a partir de un "origen" determinado.

El análisis de sobrevida, hoy en día, se desarrolla para observar y modelar el tiempo que pasa hasta la ocurrencia de diferentes clases de eventos, pero en esencia la teoría matemática-estadística es la misma. El evento puede ser la falla de una máquina, el nacimiento de un primer hijo, la tenencia de la primera casa propia, lo obtención de un grado académico, la entrada de una economía en recesión, la degradación o descomposición de un alimento, la cura de una enfermedad, etc.

Este texto ha sido elaborado pensando en un lector que demande el uso de algunas herramientas estadísticas útiles para la exploración, descripción y el análisis confirmatorio de la información, principalmente del tiempo, hasta la ocurrencia de un evento. El material estadístico que se ofrece en este texto puede ser empleado por investigadores de varias disciplinas, pues basta cambiar el escenario de los ejemplos y las ilustraciones, para hacer de este texto un instrumento de apoyo en áreas afines tales como el análisis de confiabilidad, el análisis del riesgo, el análisis de degradación, la durabilidad o longevidad y el control estadístico de calidad. El rigor matemático usado en el texto es mínimo, solo requiere de los conocimientos contenidos en un curso básico de probabilidad e inferencia estadística; no obstante, se inclu-

yen, en el apéndice A, los conceptos básicos de estas áreas para los lectores que necesiten recuperar o adquirir tales conocimientos.

En el capítulo 1 se enuncian los conceptos, definiciones y expresiones probabilísticas asociados con los datos de sobrevida. En el capítulo 2 se desarrollan las estimaciones no paramétricas de las funciones de sobrevida y de riesgo. En el capítulo 3 se presentan y caracterizan los modelos probabilísticos ligados a los datos de tiempo hasta la ocurrencia de un evento. El modelamiento estadístico de los datos bajo alguna distribución probabilística para la variable *tiempo hasta la ocurrencia de un evento*, en función de algunas covariables, lo que se conoce como modelos de regresión en datos de sobrevida, se aborda en el capítulo 5; tanto para modelos de riesgos proporcionales como de riesgos acelerados. En los capítulos 6 y 7 se trata el modelo de Cox junto con algunas de sus extensiones. El capítulo 8 trata los modelos de riesgo aditivos. El capítulo 9 contiene el análisis estadístico de datos de sobrevida desde una perspectiva bayesiana. El capítulo 10 contiene el modelamiento estadístico de la variable discreta *tiempo hasta el evento*.

El libro se debe principalmente a la bibliografía que se referencia, a mi ejercicio docente e investigativo y a las preguntas, comentarios y sugerencias de mis estudiantes y colegas investigadores del área de sobrevida. Este es un material que se puede transformar cualitativamente en la medida en que sea leído y cuestionado; por lo tanto, agradezco los comentarios y sugerencias que surjan de su estudio. Con sentimientos de gratitud y aprecio, pongo este texto a disposición de todos ustedes, amables lectores.

Luis Guillermo Díaz M.

Bogotá, 2024

Capítulo

uno

Contexto y elementos generales del análisis de sobrevivencia

$$S(t) = P(T > t) = \int_t^\infty f(u)du = 1 - F(t)$$

1.1. Introducción

El tiempo de sobrevida, el tiempo hasta la ocurrencia de un evento o los datos de duración se presentan en áreas como la medicina, la epidemiología, la biología, la demografía, la economía, la ingeniería, las ciencias actuariales y la pedagogía, entre otros campos.

La inclusión en el lenguaje estadístico de las palabras *sobrevida*, *sobrevivencia* e incluso *supervivencia* tiene su origen en la búsqueda de respuestas a las preguntas acerca de la probabilidad de que un individuo, expuesto o considerado bajo algunas condiciones o factores específicos, esté vivo después de un determinado tiempo y dentro de un periodo de estudio o seguimiento. Históricamente, estas inquietudes relacionadas con el tiempo que pasa hasta la muerte (sobrevida) de un individuo surgieron en la medicina y en áreas afines. Así, el trabajo estadístico inicial se centró en las tablas de vida humana y en la mejoría, o no, de un paciente tratado clínicamente. De esta manera, el interés estadístico se centra en modelar el tiempo de vida de una persona bajo algunas condiciones particulares. Así, el evento de interés es la muerte y se registra el tiempo inmediatamente anterior a la ocurrencia de este evento; a este tiempo se le denomina *sobrevivencia*, *sobrevida* o *supervivencia*, indistintamente. En estos vocablos el prefijo *sobre* o *super* significa *por encima* o *después de*, lo cual supone una medición del tiempo que transcurre después o por encima de un "origen" determinado.

En resumen, los términos se relacionan directamente con la ocurrencia del evento muerte, la duración del evento complementario es la sobrevida. Este es uno de los contextos históricos del origen y uso de estos términos o vocablos en la disciplina estadística. Una explicación similar se da con el término *falla*, que se refiere a la ocurrencia del evento asociado con una unidad o dispositivo mecánico (en análisis de confiabilidad).

La tarea estadística es el modelamiento, tanto probabilístico como estadístico, de la variable aleatoria *tiempo* hasta que una unidad haga la transición de un estado a otro del evento, es decir, de "no evento" a "evento". Por ejemplo, en la transición del evento "vida" a su complementario, el evento "muerte"; en la transición de "bueno" a "dañado" en el componente de una máquina; o de "soltero" a "casado" en las personas; de "consumible" a "no consumible" en los alimentos; de "no vendido" a "vendido" en un bien o servicio; de "no graduado" a "graduado" en el caso de un estudiante.

El análisis de sobrevivencia (en adelante **AS**) es una de las áreas de la estadística con mayor crecimiento en las últimas décadas. Este crecimiento se debe al desarrollo de las técnicas estadísticas y a la disponibilidad de recursos de cálculo numérico y computacionales eficientes. Como evidencia,

se destacan dos de los artículos más citados en toda la literatura estadística entre 1987 y 1989, los cuales tratan temas en esta área: el estimador de Kaplan–Meier (1958) y el modelo de Cox.

Actualmente, el interés está dirigido a modelar la variable *tiempo* desde un origen bien definido hasta la ocurrencia de algún evento, **E**. Es decir, la transición del estado "no evento", $\mathbf{E}^c$, al estado "evento", **E**, dentro de un periodo de estudio u observación particular sobre un objeto o unidad específica. Por ejemplo, la transición del evento sano a enfermo, o recíprocamente, de enfermo a sano, si se quiere observar la duración de una enfermedad.

Con esto, se ha extendido el estudio de la variable aleatoria *tiempo* hasta la ocurrencia de eventos diferentes a la muerte (o a la falla), y por lo tanto, se ha extendido al estudio y aplicación en otras áreas del conocimiento. Así, por ejemplo, en la ingeniería este tipo de estudios relacionados con el tiempo que pasa hasta el evento "falla" de un dispositivo o mecanismo se denomina análisis de *confiabilidad*; en las ciencias sociales, el interés puede estar dirigido a modelar el tiempo hasta la ocurrencia del evento "tener el primer hijo", "adquirir la primera vivienda propia", "obtener la graduación como profesional", "contraer el primer matrimonio"; en economía, puede ser el tiempo que pasa hasta que la economía de un país entra en recesión, entre otras situaciones que hacen apropiado el término *evento* y que no siempre corresponden a una falla. No obstante, en este texto se hace uso de estos términos de manera equivalente.

Generalmente, la variable respuesta en el análisis de sobrevivencia es el tiempo que pasa hasta la ocurrencia de un evento, **E**, de interés. Este tiempo se conoce como *tiempo hasta la ocurrencia del evento*, en adelante, *tiempo hasta el evento*. A continuación, se mencionan algunos tiempos de interés asociados con un evento.

1. Tiempo hasta la muerte de una persona o cualquier ser vivo.
2. Tiempo desde el diagnóstico de una enfermedad hasta la muerte del paciente.
3. Tiempo desde el diagnóstico de una enfermedad hasta la cura de la enfermedad.
4. Tiempo transcurrido entre dos averías de un componente o elemento de una máquina.
5. Tiempo hasta que una mujer tiene su primer hijo.
6. Tiempo hasta que un estudiante se gradúa de un programa académico.

7. Tiempo hasta que un niño o niña aprende a leer (o a escribir).

8. Tiempo hasta que una persona, con licencia de conducción, tiene su primer choque automovilístico.

9. Tiempo hasta que un hijo(a) deja su núcleo familiar (emancipación).

10. Tiempo hasta cuando una familia adquiere su primera vivienda propia.

11. Tiempo hasta que un producto, puesto en un punto de exhibición y venta, es comprado por un consumidor.

12. Recorrido (tiempo o kilometraje) hasta que un neumático de un automóvil sufre una perforación ("pinchazo").

Así, la variable aleatoria *tiempo hasta el evento*, **T**, se asume como una variable de tipo discreto o continuo; no obstante, la teoría y aplicación se han desarrollado principalmente sobre la consideración continua de la variable tiempo. Además, la unidad permanece (antes del evento) de manera continua en el tiempo en un estado de no evento, $\mathbf{E}^c$. Otra cosa es que su medición se haga de manera discreta. Además, en algunos casos, este tiempo no se registra de manera permanente, es decir, con un cronometraje continuo, sino que se mide el tiempo de permanencia en funcionamiento de un aparato. Así, el tiempo calendario y el tiempo en funcionamiento no coinciden, pues pueden haber periodos de estado "apagado" o de inactividad del aparato o dispositivo.

Aunque la variable *tiempo hasta la ocurrencia de un evento* es la variable aleatoria central para el análisis estadístico de sobrevida, hay casos en los cuales la variable *tiempo* no es la que se observa o la que interesa directamente, sino otra característica asociada con el evento de interés **E**. Por ejemplo, en el evento perforación o "pinchazo" de un neumático (último caso) la variable a considerar puede ser el desgaste o profundidad de la superficie de contacto con el piso (banda de rodamiento) de la llanta. Un caso similar se tiene con la variable *kilómetros o millas de recorrido* en el caso del motor de un automóvil, la turbina de un avión o un barco.

En general, la literatura de los métodos estadísticos y de los modelos propuestos para el análisis de datos de sobrevida asume que el tiempo hasta el evento de interés es una variable aleatoria continua de soporte positivo. No obstante, en algunas ocasiones, el envejecimiento o vetustez de los elementos (por ejemplo, componentes, sistemas, subsistemas, especímenes, estructuras, órganos vitales, unidades) no se mide en términos temporales

o cronológicos. En tales ocasiones, la vida útil se mide por medio de otras variables aleatorias. Por ejemplo, la cantidad de kilómetros recorridos, la resistencia de las muestras de material hasta su ruptura, el nivel de degradación, la flexibilidad de un adhesivo o el número de ciclos hasta que una muestra de material falla a causa de la fatiga. Además, la terminología *variable aleatoria de tiempo al evento*, que se denota en lo sucesivo por **T**, se usa ampliamente para hacer referencia a cualquier variable aleatoria continua positiva. No obstante la naturaleza intrínsecamente continua de la variable aleatoria *tiempo*, debido a las circunstancias de tipo experimental o a los procedimientos de medición u observación sobre las unidades, se debe optar por considerar la variable tiempo de manera discreta o discretizada. Cualquier modelo probabilístico asociado con la variable aleatoria *tiempo hasta el evento* a menudo se denomina una *distribución de vida*.

Una característica casi exclusiva de los datos de sobrevivencia es que el registro o evidencia del tiempo para el evento en algunas unidades no se conoce con exactitud, o tan solo se sabe que se presenta dentro de un cierto periodo de tiempo; a esto se le denomina *censura*. Así, si a una unidad le ocurre el evento después del periodo de estudio, se dice que este tiempo tiene *censura a derecha*. La *censura a izquierda* ocurre cuando tan solo se conoce que el individuo ha experimentado el evento de interés antes de empezar el estudio. Y la *censura por intervalo* se presenta cuando la única información disponible es que el evento ocurrió dentro de un intervalo específico de tiempo.

Otra característica de los datos de sobrevida se produce cuando solo se consideran o reconocen para el análisis, por decisión o a criterio del investigador, los tiempos de las unidades o individuos para los cuales el evento ocurre dentro de una ventana o intervalo de observación determinado. Esto se conoce como *truncamiento*.

La presencia de censura hace que se tenga una observación parcial o incompleta de la variable respuesta *tiempo*. Por ejemplo, si se hace seguimiento a un paciente y este se muda de ciudad, muere por causa diferente a la estudiada, termina el estudio o se interrumpe su acompañamiento (en inglés *drop out*), entonces la información disponible es que el tiempo al evento (muerte) es diferente (superior o inferior) al tiempo especificado para la terminación del estudio u observación. La mayoría de las técnicas estadísticas clásicas (como el análisis de varianza o los métodos de regresión clásica) tienen limitaciones para aplicarse cuando hay presencia de información censurada porque se requieren de manera explícita los tiempos exactos al evento. Se hace entonces necesario el uso de métodos de análisis estadísticos que consideren e incorporen la información contenida en los datos, tal como

la censura y el truncamiento. Cuando no hay censura ni truncamiento, el conjunto de tiempos se denomina *completo*.

El análisis de sobrevivencia se refiere básicamente a situaciones que involucran algunos datos de tiempo con censura y, en algunos casos, están truncados. Esto pasa en las ciencias de la salud, las ciencias sociales o la ingeniería, donde se requiere estimar algunas características relacionadas con la distribución de la variable *tiempo*, tales como el tiempo medio de falla, la esperanza de vida, los percentiles (mediana, cuartiles, quintiles) de la variable tiempo, la vida media residual o la probabilidad de que el evento ocurra después de cierto tiempo.

En el análisis de confiabilidad se quiere determinar la probabilidad de que un producto dure más de un tiempo determinado, con lo cual se podría establecer un periodo de garantía para el comprador sobre el producto. En criminalística interesa el tiempo entre la liberación de un reo y la ocurrencia de un nuevo delito cometido por este. En estudios socioeconómicos, por ejemplo, el interés se centra en observar, sobre una persona, el evento "empleado" o "desempleado", los cambios de empleo, el tiempo que dura alguien en estado de desempleo, o el tiempo para obtener una jubilación. En demografía pueden interesar los eventos como el nacimiento, la muerte, el matrimonio, el divorcio o la migración.

Para cada una de las situaciones o ejemplos señalados se puede estar interesado en estudiar el efecto o la asociación de la ocurrencia de un evento, o su ausencia, bajo la consideración de algunas covariables o factores asociables al mismo. Es decir, se trata entonces de considerar estadísticamente la variable *tiempo hasta el evento*, $\mathbf{T}$, en relación con otras covariables contenidas en un vector X, las cuales pueden ser categóricas o numéricas, dependientes del tiempo o no.

Algunas veces es de interés solamente la distribución de los tiempos hasta el evento de un único grupo. Frecuentemente, el interés se dirige a comparar los tiempos hasta el evento en dos o más grupos para evidenciar, por ejemplo, si los tiempos al evento de los individuos son mayores sistemáticamente en un grupo respecto de otro(s). Alternativamente, los valores de las variables explicativas o covariables deben estar disponibles para cada individuo. Estas variables se consideran que están asociadas con la sobrevivencia. Así, por ejemplo, el tiempo de vida de una máquina puede estar influenciado por el esfuerzo ejercido sobre esta, el material del cual está hecha, las sustancias con que tenga contacto o la temperatura del área de trabajo donde funciona. Por lo tanto, estas condiciones pueden tomar el papel de variables explicativas en la sobrevivencia de la máquina objeto de estudio. En investigación clínica, es común que de forma rutinaria se colecte una gran cantidad de

información (covariables) sobre cada paciente; el objetivo de estos estudios es determinar y resumir el efecto conjunto de estas variables explicativas o covariables de la sobrevida del paciente.

1.2. Estructura de los datos de sobrevida

Los datos de sobrevivencia se caracterizan por el registro del tiempo hasta la ocurrencia del evento y, frecuentemente, por las censuras, el abandono (*drop out*) o el truncamiento. Estos componentes definen y estructuran la variable respuesta *tiempo hasta la ocurrencia del evento*. En la mayoría de los estudios de sobrevida se incluyen otras variables, llamadas covariables, las cuales se observan y registran sobre la unidad a la cual se hace seguimiento respecto de la ocurrencia o no del evento de interés. Los siguientes elementos constituyen y definen el tiempo hasta el evento:

- La unidad sobre la cual se registra el evento.
- El evento de interés.
- El tiempo origen o inicial y el tiempo final del estudio.
- La escala de medida del tiempo hasta que ocurre el evento.
- La distribución probabilística de la variable *tiempo* hasta que ocurre el evento.
- El tipo de censura y el truncamiento del tiempo.
- Las covariables, factores o condiciones a los cuales se les pueden atribuir la ocurrencia o no del evento.

Estos elementos deben ser definidos claramente y contextualizados con el marco conceptual del estudio (ciencias de la salud, sociales, ingeniería, entre otras). A continuación se definen estos conceptos.

La *unidad* corresponde al objeto o individuo sobre el cual se registra el evento. La unidad puede estar constituida por varias unidades como una familia, una camada de animales, una parcela de plantas o un sistema de componentes de una máquina; en otros casos la unidad corresponde a objetos individuales como una persona, un animal, una planta o un componente mecánico. Estas unidades pueden corresponder a unidades experimentales o muestrales, unidades de estudio sobre una cohorte o estudios de casos.

Otra situación que puede presentarse es cuando las unidades se agrupan, de manera jerárquica, en conglomerados de unidades. Por ejemplo,

pacientes tratados en el mismo hospital, automóviles fabricados en la misma planta, estudiantes que asisten a la misma escuela; a la vez, cada uno de estos grupos o conglomerados puede formar parte de un grupo más grande y así sucesivamente. Para el caso de los estudiantes, estos pertenecen a un grado o nivel escolar que está dentro de una escuela; esta escuela está incluida en una ciudad; la ciudad hace parte de algún departamento o región, y así hasta cubrir todo el sistema educativo. Este tipo de estructura de las unidades y de los datos se denomina *multinivel*, y cada uno de los niveles corresponde a una aglomeración de unidades. En el modelamiento y análisis de estos datos se debe considerar e incluir las características observadas en cada nivel. Para los ejemplos anteriores se deben considerar los efectos de los factores asociados con el hospital, la planta o la escuela, respectivamente, sobre el tiempo hasta el evento.

El *evento* debe pertenecer a un conjunto de resultados posibles (espacio muestral) del estudio observacional, el cual puede presentarse, sobre cada unidad, una o varias veces. El evento debe estar definido de forma no ambigua. En el caso de muerte o falla parece claro que no habría duda sobre la ocurrencia o no del evento, pero en el evento "el producto es apto para el consumo", que se da con alimentos o fármacos, deben establecerse las condiciones físicas, químicas y bioquímicas para considerar que el momento de consumo de un producto es apropiado para la salud del consumidor.

Como el evento ocurre cuando la unidad cambia de estado, de $\mathbf{E}^c$ a $\mathbf{E}$ (sano a enfermo, bueno a dañado, en existencia a vendido, no graduado a graduado, soltero a casado, entre otros), se debe advertir que el tiempo cuenta hasta el cambio inmediato o instantáneo de estado de la unidad. Es decir, no se considera el tiempo que la unidad permanece en el estado o el evento. Por ejemplo, la duración de una enfermedad o la permanencia en estado averiado de un aparato mecánico.

En algunas aplicaciones hay poca o ninguna arbitrariedad en la definición del evento. En otras, por ejemplo, en algunos contextos industriales, la falla se define como el primer momento en el cual el desempeño, medido de alguna forma cuantitativa, cae por debajo de un nivel (umbral) técnicamente aceptable y previamente establecido.

Un estudio de tiempo para el evento de cinco unidades es ilustrado en la figura 1.1, en la cual se muestra que el ingreso al estudio de las unidades no se da en el mismo tiempo. El tiempo de entrada al estudio se representa por "►". El intervalo 0 a t_0 se conoce como tiempo de alistamiento o reclutamiento, durante el cual se acopian o reúnen las unidades del estudio. El tiempo t_0 es establecido por el investigador y corresponde al origen de

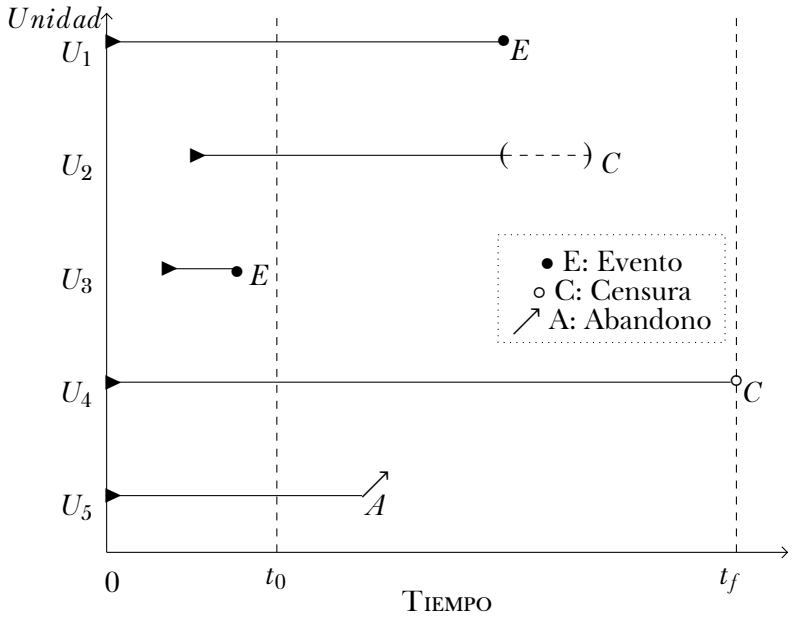

Figura 1.1. Tiempo al evento en estudio sobre cinco unidades.

medición del tiempo hasta el evento. De la misma manera el investigador establece el tiempo final, t_f, del estudio.

Sobre la unidad U_1 el evento ocurre durante el periodo de estudio. Sobre la unidad U_2 el evento ocurre pero no hay precisión del momento exacto de ocurrencia; solo se sabe que ocurre dentro de un intervalo de tiempo (censura por intervalo). En la unidad U_3 el evento ocurre antes del tiempo de inicio del estudio; o dicho de otra manera, cuando se estableció el tiempo de inicio del estudio, t_0, el evento ya le había ocurrido a esta unidad (esto se denomina censura a izquierda). En la unidad U_4, se termina el tiempo considerado como fin del estudio y el evento no se ha observado sobre ella; tal vez ocurra después de t_f (censura a derecha). Finalmente, la unidad U_5 se pierde o abandona el estudio, en consecuencia, no se sabe si le ha ocurrido el evento; a este tipo de unidades se les considera como abandonos (en inglés, *drop outs*). Y aunque se ha introducido el término *censura* para decir que no se sabe con precisión el momento de ocurrencia del evento, este se define y caracteriza en la sección 1.2.2.

Es deseable que, bajo ciertas restricciones de algunas variables explicativas, todos las unidades del estudio sean tan comparables como sea posible en sus tiempos de origen. El tiempo de origen no necesita ser y usualmente no está en el mismo tiempo calendario para cada individuo; cada unidad tiene

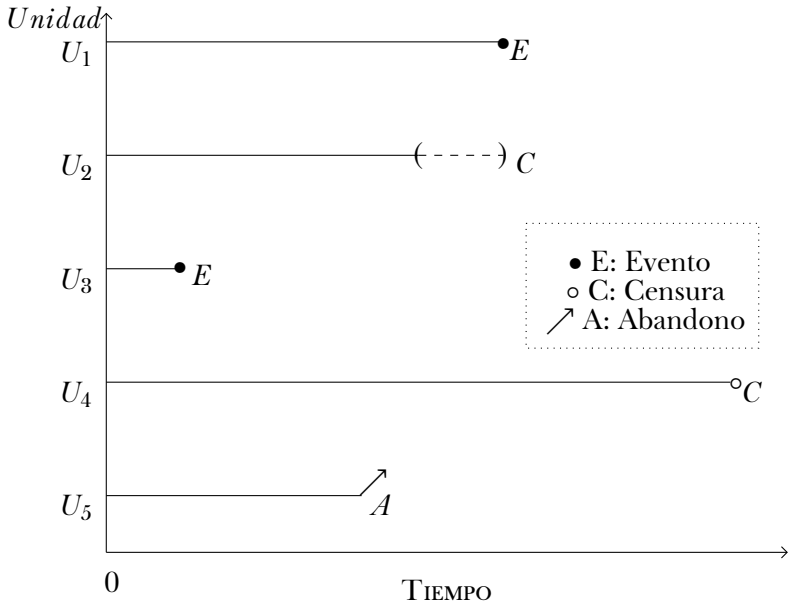

Figura 1.2. Tiempo al evento en cinco unidades con origen común.

su propio tiempo origen. Se puede optar por hacer una traslación del tiempo origen, donde, nominalmente, todas las unidades tengan un mismo tiempo origen. De esta forma, el periodo de tiempo desde este origen común hasta el evento es el tiempo al evento (tiempo de sobrevida). Los tiempos de las unidades sobre las que no se evidencie el evento se consideran censuradas.

En estudios experimentales, generalmente el origen puede ser definido por el investigador que diseña y desarrolla el experimento; no obstante, en muchos estudios el tiempo de origen es propio para cada unidad o no se tiene con precisión. Una representación conveniente de estos datos es trasladar el origen de cada unidad a un origen común 0 para todas las unidades mostradas en la figura 1.2.

Otra representación de este tipo de datos es conocida como los diagramas de Lexis. En estos, sobre el eje horizontal se representa el tiempo calendario, la "longevidad" de la unidad es representada sobre una línea de 45^o, y el tiempo que la unidad permanece en el estudio se representa por la altura del rayo de la duración o longevidad en el eje vertical. En la figura 1.3, sobre las unidades U_1 y U_3 el evento **E** ocurrió antes de finalizar el estudio, en consecuencia, se conoce el tiempo exacto al evento. Las unidades U_{2cd} y U_{4cd} posiblemente experimentan el evento después que el finaliza el estudio; este tipo de censura se conoce como censura a derecha *Tipo I generalizada*.

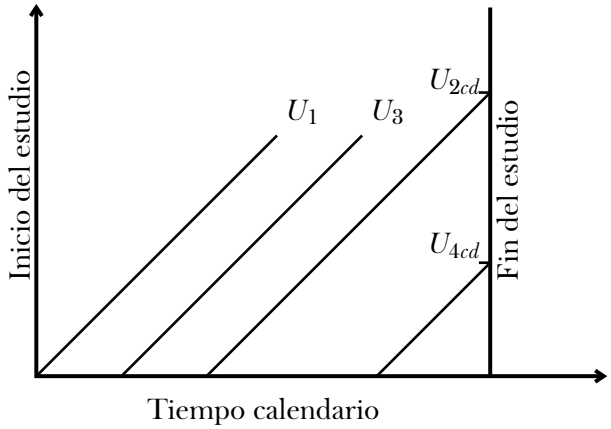

Figura 1.3. Diagrama de Lexis para censura generalizada Tipo I.

El tiempo de inicio u origen puede estar asociado con el momento de inicio del estudio o con la ocurrencia de un evento relacionado con el evento de interés **E**; no obstante, en muchos estudios de tipo experimental, este es determinado por condiciones bajo las cuales se desarrolla la observación, las que pueden ser de tipo logístico o por condicionamientos físicos, naturales, administrativos, de costos o sociales. Algunas consideraciones similares a las del tiempo de origen se tienen para el tiempo de finalización del estudio.

La *escala de medición* generalmente es en tiempo real (calendario), aunque existen otras alternativas, como la distancia recorrida (kilometraje) para automóviles, los ciclos o el recorrido en millas para turbinas de avión, el desgaste (en mm o micras) de un componente de una máquina, o los intentos hasta obtener un resultado "exitoso" en el caso de la exploración de pozos petroleros productivos.

La *distribución probabilística* de la variable aleatoria *tiempo hasta la ocurrencia del evento* se debe establecer si se conoce el modelo probabilístico que genera los datos. El modelo probabilístico se asume por estudios previos o semejantes sobre la variable *tiempo hasta la ocurrencia del evento*, es decir, por la exploración y descripción de los datos asociados con la variable, pues gráficos tales como histogramas, función de sobrevida empírica, diagramas de tallos y hojas diagramas de cajas, estadísticas evaluadas en datos empíricos o preliminares como la media, la mediana, percentiles, cuantiles, varianza, coeficientes de variación, entre otros, pueden advertir sobre el tipo de distribución ajustable al escenario conceptual de la variable. Una característica común a todas las variables *tiempo hasta la ocurrencia del evento* es que toman valores en los números reales positivos.

Un aspecto adicional son *las covariables*, que son factores o condiciones que posiblemente son asociables a la ocurrencia o no del evento, pues la realización de este puede deberse o atribuirse a unos factores dentro de una multiplicidad de ellos. Esto se da, por ejemplo, en situaciones de fallo por causas múltiples, como es el caso de la muerte por la concurrencia de fallas en distintos órganos o el desperfecto de una máquina cuyos fallos son atribuibles al sistema de refrigeración, el sistema eléctrico o la carga. Estos casos se conocen en la literatura del análisis de sobrevida como *riesgos en competencia*.

1.2.1. Variable aleatoria *tiempo hasta el evento*

Como se describe en la sección anterior, el tiempo de origen del estudio debe estar claramente definido para cada unidad. Este no necesariamente debe coincidir con el tiempo calendario y, en este sentido, puede no ser el mismo para todas las unidades. Así, por ejemplo, el tiempo origen puede ser el momento en que a un individuo se le diagnostica una enfermedad, caso en el cual el origen puede ser la edad del individuo al momento del diagnóstico o el momento en el que inicia un tratamiento para la enfermedad. En algunos estudios experimentales, el tiempo origen puede definirse sincrónicamente para todas las unidades.

De igual forma, el *tiempo de finalización del estudio* puede ser determinado por el investigador. Este tiempo puede estar condicionado a las políticas en el campo de investigación, la logística disponible para el estudio, la naturaleza del evento a observarse o a la necesidad de tomar decisiones sobre el área de estudio. En el caso de un estudio epidemiológico, el tiempo de finalización del estudio esta sujeto tanto a la naturaleza de la enfermedad en consideración como al tiempo para tomar una decisión en salud pública sobre el grupo o población humana, animal o vegetal que es objeto de estudio.

Como una formalización de los casos específicos de tiempos para un evento presentado, se puede declarar que una *variable aleatoria* (*v. a.*) $\mathbf{T}$ es una *v. a.* de sobrevivencia si una respuesta observada t de $\mathbf{T}$ está definida en el intervalo $[0, +\infty)$, donde t es el tiempo para el evento $\mathbf{E}$. De otra manera equivalente, $\mathbf{T}$ mide el tiempo de permanencia de la unidad en el estado complementario, $\mathbf{E}^c$. La figura 1.4 ilustra esta variable aleatoria.

Como se puede apreciar, la *v. a.* $\mathbf{T}$ es de naturaleza continua con función de densidad $f(t)$, pero en algunas situaciones, por razones logísticas o de medición del tiempo, esta debe considerarse como discreta (véase el ejemplo 7, presentado en la sección 1.2). Así, la *v. a.* $\mathbf{T}$ toma los valores t_j, con

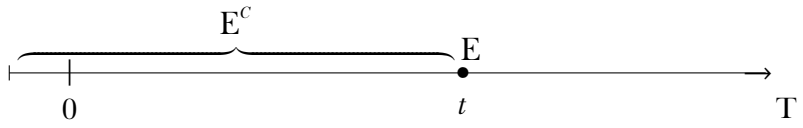

Figura 1.4. Línea de tiempo.

$j = 1, 2, \ldots$, con función de probabilidad $p(t_j) = P(T = t_j)$, $j = 1, 2, \ldots$. Se asume que $t_1 < t_2 < \cdots$. En adelante, se hace referencia a una función de probabilidad para el caso discreto o continuo indistintamente; se precisará que el caso es continuo o discreto cuando sea necesario. La probabilidad de que el evento ocurra sobre una unidad específica, antes de un tiempo t, para el caso continuo, viene dada por:

$$P(T \leq t) = \int_0^t f(u)du) = F(t),$$

la cual se conoce como la función de distribución acumulada o simplemente la función de distribución, $F(t)$, de la variable aleatoria $\mathbf{T}$.

Para el caso discreto,

$$P(T \leq t) = \sum_{t_j \leq t} p(t_j) = F(t).$$

Se advierte que para el caso continuo $P(\mathbf{T} \leq t) = P(\mathbf{T} < t)$.

1.2.2. Censura y truncamiento

Los estudios longitudinales (indexados en el tiempo) son de dos tipos: los *prospectivos* y los *retrospectivos*. Un estudio *prospectivo* es aquel en el cual la medición de la exposición (covariables y factores asociables con la ocurrencia del evento) se realiza en una unidad antes de la ocurrencia del evento y, por tanto, antes de su registro. Es decir, el inicio del estudio es previo al evento de interés y los datos se consiguen a medida que van ocurriendo. Por el contrario, en un estudio *retrospectivo*, la medición de la exposición se produce después de que se ha evidenciado la ocurrencia del evento sobre una determinada unidad. De esta manera, el inicio del estudio es posterior a los hechos estudiados y los datos se recogen de archivos o de entrevistas sobre hechos sucedidos (por ejemplo, historias de vida).

Una característica de los estudios de sobrevida es la presencia de observaciones incompletas o parciales. Estas observaciones se denominan *censuradas* y pueden ocurrir por múltiples situaciones, tales como la pérdida del

seguimiento a la unidad, porque se abandona el estudio o porque no se tiene precisión acerca del momento en que el evento ocurre.

De forma general, la censura ocurre cuando se sabe que algunos tiempos para el evento han ocurrido solamente dentro de ciertos intervalos y el resto de los tiempos se conocen exactamente. Como ya se anotó, existen varias categorías de censura, principalmente, censura por la derecha, censura por la izquierda y censura por intervalo. Para identificar adecuadamente el tipo de censura que presentan los datos de tiempo para un evento, se debe considerar la forma como se han registrado. Cada tipo de censura puede corresponder a una función diferente de verosimilitud, la cual suele ser la base para la inferencia en el modelamiento estadístico.

Así, la información disponible sobre estas unidades es que el tiempo para la ocurrencia del evento no se ha observado de manera explícita (censura), pues puede ser mayor al tiempo de seguimiento, menor al tiempo de inicio del estudio o estar dentro de un intervalo de tiempo. Adicionalmente, el protocolo del estudio puede restringir el tiempo para el evento si este se presenta después o antes de unos tiempos establecidos (truncamiento).

Para la estimación de parámetros y el análisis estadístico es necesario incorporar todas las piezas de información probabilística, sean censuradas o no, truncadas o no. La justificación es que la información censurada o truncada, aunque incompleta, aporta a la estimación de parámetros y al análisis de la variable tiempo para el evento; su omisión o exclusión del análisis estadístico puede sesgar las conclusiones del estudio.

Además, para el análisis estadístico de datos con presencia de censura, se debe considerar el supuesto de que el tiempo hasta la ocurrencia del evento de una unidad es independiente de cualquier mecanismo que cause que ese mismo tiempo resulte censurado. Un ejemplo del incumplimiento de este supuesto se tiene en tratamientos médicos, en los cuales un individuo se retira del estudio porque considera que su estado de salud ha mejorado (debido al tratamiento) o porque, por razones bioéticas, debe retirarse del estudio (deterioro de la salud). Cuando se cumple este supuesto de independencia, se dice que el proceso de censura es *no informativo*. El no cumplimiento de este supuesto puede invalidar los resultados del análisis estadístico en términos de la calidad de las estimaciones y de la validez de la inferencia estadística que desde los datos se pretenda hacer.

La *censura Tipo I* se tiene cuando el evento es observado solo si ocurre antes de algún tiempo preespecificado (generalmente, tiempo de estudio).

La *censura Tipo II* corresponde al caso en el cual el estudio se desarrolla hasta que el evento haya ocurrido sobre r individuos , donde r es algún número entero predeterminado ($r \leq n$); así se dispone de r tiempos al evento

observados y $(n - r)$ tiempos censurados. Esta censura se presenta generalmente en estudios o ensayos sobre la duración (confiabilidad) de equipos. En este caso, el tiempo de inicio del estudio debe ser el mismo para todas las unidades.

Otra característica de los estudios de sobrevida es el *truncamiento*. El truncamiento de los datos de sobrevida se presenta cuando solo se consideran (por decisión del investigador) aquellos individuos o unidades cuyos tiempos para el evento están dentro del intervalo o ventana de observación (T_i, T_d). La información de un individuo cuyo tiempo para el evento no está dentro de este intervalo no es considerada ni incluida en el análisis estadístico, situación que no ocurre con la censura.

Un caso especial de este tipo de truncamiento consiste en considerar únicamente alguno de los valores T_i o T_d. Así, con T_i se consideran los tiempos para el evento tal que $T > T_i$, a este tipo de datos se les denominan *truncados por la izquierda*. Por ejemplo, un parangón sería el caso de las partículas que se observan a través de un microscopio, donde solo se consideran aquellas cuyo diámetro sea superior a un valor preestablecido. De manera análoga, se tienen los datos *truncados a derecha*, los cuales corresponden a tiempos para el evento menores que un valor T_d. Hay una similitud con la observación de cuerpos celestes mediante un telescopio, pues ahí se consideran solo aquellos cuya distancia sea menor que un determinado valor. Los órganos de los sentidos (en el caso de los humanos y de algunos animales) tienen truncamiento de intervalo, pues perciben ciertas características (tamaño, color, olor, sonido) si estas están dentro de cierto intervalo (umbral mínimo y máximo).

1.2.3. Censura a la derecha Tipo I

Se asume que el estudio se desarrolla y observa dentro de un periodo de tiempo $[t_0, t_f]$. Si el evento es observado y ocurre en un tiempo menor o igual que t_f $(\mathbf{T} \leq t_f)$ se dice que el tiempo es exacto, si el evento posiblemente ocurre después del tiempo t_f, el tiempo no es exacto y se denomina *censurado a derecha* $(\mathbf{T} > t_f)$.

Para el caso del evento "primer hijo" de una mujer, puede que se termine el tiempo de estudio (por ejemplo, de los 18 a los 40 años) y la mujer no haya tenido su primer hijo. Para el evento "tiempo hasta aprender a leer o escribir" en un niño, cuya edad está entre 4 a 6 años, el estudiante puede que muestre este aprendizaje después de los 6 años. En una máquina, puede que se termine el tiempo de observación y esta siga funcionando adecuadamente; tal vez el desperfecto le ocurra después del tiempo final del estudio.

Se asume que para un individuo hay un tiempo para el evento $\mathbf{T}$ y un tipo de censura fijo C_d. Se asume que los $\mathbf{T}$ son independientes e idénticamente distribuidos con $fdp\ f(t)$ y función de sobrevida $S(t)$. El tiempo para el evento exacto de un individuo es conocido si $\mathbf{T} < C_d$; si $T \geq C_d$, la censura ocurre en C_d. Así, los datos son de la forma (T, δ), donde $\delta = 1$ si T corresponde al tiempo para el evento y $\delta = 0$ si hay censura; es decir,

$$\delta_i = \begin{cases} 1, & \text{si en la unidad } i \text{ se observa el evento,} \\ 0, & \text{si en la unidad } i \text{ no se observa el evento.} \end{cases}$$

En caso de censura a la derecha se asume como tiempo para el evento a $\widetilde{\mathbf{T}} = \text{mín}\{\mathbf{T}, C_d\}$; es decir, el menor de los tiempos $\mathbf{T}$ o C_d; información que debe ser incorporada al análisis o modelamiento estadístico de estos datos.

1.2.4. Censura a la derecha Tipo II

Se presenta en estudios en los cuales se continúa hasta cuando se tenga el tiempo para el evento de los primeros r individuos ($r \leq n$, r es un valor fijo). En consecuencia, se tienen r tiempos observados para el evento y $n - r$ censuras.

Un ejemplo de censura Tipo II es el caso de observar un dispositivo electrónico hasta que se dañe. Se inspeccionan dispositivos de un lote de producción hasta que se observen los tiempos de duración de 10 de estos; por supuesto, habrá que observar un número $n \geq 10$ de tales dispositivos. Esto implica consideraciones de costos, pues no será procedente tener que probar un número elevado de dispositivos para encontrar los 10 datos requeridos.

1.2.5. Censura a la derecha bajo riesgos en competencia

En algunos casos (experimentos o estudios) se tiene la situación de *riesgos competentes*. Corresponden al hecho de que algunas unidades pueden mostrar el evento $\mathbf{E}$ (generalmente falla o muerte) debido a una de varias causas, llamadas *riesgos competentes*. Un ejemplo son las causa específicas de mortalidad humana, tales como enfermedades cardíacas, renales, cáncer, etc. Otro ejemplo es el evento "daño o desperfecto en un motor", pues esta situación puede presentarse por problemas en el componente eléctrico, en el sistema de refrigeración, en el sistema de lubricación o en el sistema de alimentación (combustible).

El interés puede dirigirse a estimar una distribución marginal para un evento, pero algunas unidades pueden experimentar un riesgo competente

que provoca que tales unidades deban removerse del estudio. De esta forma, el evento de interés no es observado en las unidades que experimentan el evento competente y estas quedan con censura aleatoria a la derecha en ese tiempo.

Cuando el interés está en modelar el riesgo a un evento particular **E** pero algunas unidades bajo estudio pueden experimentar algún evento "competente", $\mathbf{E_k}$, esto causa que la unidad sea retirada o removida del estudio. En tales casos, el evento de interés no es observable para las unidades que experimentan el evento competente y estas unidades quedan censuradas aleatoriamente a la derecha en ese tiempo (momento de ocurrencia del evento competente).

1.2.6. Censura a la izquierda

La censura a la izquierda se presenta cuando tan solo se conoce que el evento ha ocurrido antes del tiempo de inicio de la observación. Así, un tiempo hasta el evento **T**, asociado con una unidad específica del estudio, es considerado *censurado a izquierda* si el evento de interés, **E**, ha ocurrido antes que la unidad haya sido observada por el investigador al tiempo C_i. Para tales unidades, se asume que han tenido el evento en algún tiempo **T** anterior al tiempo C_i.

En caso de censura a la izquierda, se asume como tiempo para el evento a $\widetilde{T} = \text{máx}\{T, C_i\}$; es decir, el mayor de los tiempos T o C_i; el cual debe ser considerado e incorporado al análisis o modelamiento estadístico.

Algunas veces, si la censura por la izquierda ocurre en el estudio, la censura por la derecha puede ocurrir también y los tiempos al evento son considerados doblemente censurados. Así, los datos pueden ser representados por una pareja de variables (T, δ) donde $T = \text{máx}\{\text{mín}\{X, C_d\}, C_i\}$ es el tiempo de estudio y δ es una variable indicadora definida como sigue:

$$\delta_i \begin{cases} 1, & \text{si } T \text{ es el tiempo hasta la ocurrencia del evento} \\ 0, & \text{si } T \text{ es el tiempo censurado a derecha} \\ -1, & \text{si } T \text{ es el tiempo censurado a izquierda.} \end{cases}$$

Un ejemplo es cuando se pregunta a una persona sobre el consumo de algún fármaco por primera vez. La respuesta puede ser que no se acuerda cuándo fue la primera vez. Aquí el tiempo de censura a izquierda, C_i, es la edad actual de la persona. La censura a la derecha y a la izquierda pueden presentarse simultáneamente en el estudio. Otro ejemplo es el caso del evento "aprender a leer o escribir". Puede ocurrir que, al observar a un niño en re-

lación con esas habilidades y destrezas (lectoescritura), este ya las haya desarrollado cuando ingresa al grado escolar donde se espera que las aprenda (primero de primaria, en el sistema educativo colombiano); así, para estos niños la variable *tiempo* para la lectoescritura estaría censurada a izquierda.

En estudios prospectivos, algunos eventos como la muerte no tiene sentido que tengan censura a izquierda.

1.2.7. Censura por intervalo

En esta situación solo se sabe que el tiempo al evento **E** es mayor que el último tiempo de observación en el que este no ha ocurrido (la unidad permanece en el estado $\mathbf{E^C}$) y es menor o igual que el primer tiempo de observación en el que se ha visto el cambio (la unidad ahora se muestra en el estado **E**). Así se dispone de un intervalo que contiene el tiempo real (pero no observado) de ocurrencia del evento o cambio del estado $\mathbf{E^C}$ a **E**. En consecuencia, los datos del tiempo para el evento censurados a la derecha o a la izquierda son un caso especial de datos censurados por intervalos, donde los intervalos tienen un extremo no finito, pues son de la forma $(C_d, +\infty)$ y $(-\infty, C_i)$, respectivamente.

Este tipo de censura se presenta cuando se tiene un estudio longitudinal donde el seguimiento del estado de las unidades se realiza periódicamente, y por tanto, el evento solo puede conocerse entre dos periodos de revisión, generando un intervalo de la forma (C_1, C_2) para cada unidad en el estudio. Se advierte que esta notación no indica que los extremos del intervalo de censura coincidan con los puntos para la censura a izquierda y a derecha, respectivamente.

Un ejemplo de datos censurados por intervalo surge en el estudio del virus de inmunodeficiencia humana (VIH). En pacientes seropositvos se puede presentar el evento o episodio del síndrome de inmunodeficiencia adquirida o SIDA (pacientes cuya carga viral está por debajo de cierto umbral). En estos casos, la determinación del inicio del SIDA suele basarse en análisis de sangre hechos en laboratorios, los cuales, obviamente, solo pueden realizarse de manera periódica pero no continua. Como consecuencia, solo se dispondría de datos con censura por intervalo para el diagnóstico del tiempo o momento del SIDA.

Un caso similar se da en los estudios sobre los tiempos de infección del VIH. Si un paciente es seropositivo al comienzo de un estudio, entonces el tiempo o momento en el que se contrajo la infección del VIH suele determinarse mediante un análisis retrospectivo del historial médico del enfermo. Por lo tanto, solo se puede obtener un intervalo cuyo extremo izquierdo es

definido por la última fecha de prueba en suero sanguíneo con resultado negativo para el VIH y cuyo extremo derecho es la primera fecha de prueba en suero sanguíneo del resultado positivo para el tiempo de infección del VIH.

Para el diagnóstico de un embarazo (el evento con precisión sería la fecundación de un óvulo), los médicos o ginecólogos lo ubican, de manera general, como el intervalo entre la última menstruación y la amenorrea (ausencia de menstruación) junto con otra sintomatología determinada médicamente. Este puede ser otro ejemplo de censura por intervalo.

Para el caso "tiempo al desperfecto de una máquina" puede que este se presente al momento de encender o hacer uso de esta. Así, el tiempo para el evento "falla" se establecería entre el momento que se detecta su no funcionalidad y el tiempo de uso inmediatamente anterior.

En el seguimiento a niños para establecer el evento "aprendizaje de lectura o escritura", el docente o maestro que acompaña a estos estudiantes generalmente solo puede dar cuenta de esto entre dos periodos consecutivos de evaluación de estas habilidades. Es decir, una evaluación donde el estudiante no muestra estas habilidades y otra evaluación consecutiva en la que sí las evidencia.

1.3. Censura aleatoria (Tipo III)

En algunos estudios, el período de observación es fijo y las unidades ingresan al estudio en diferentes momentos durante ese período; por ejemplo, los pacientes que padecen una enfermedad o las maquinas que tienen alguna avería. Algunos pueden tener el evento (**E**) de interés antes de finalizar el estudio; por lo tanto, se conocen sus tiempos hasta el evento. Otras unidades pueden retirarse antes de terminar el estudio y perderse durante el seguimiento, mientras que otros pueden no haber tenido el evento al final del estudio. Para las unidades "perdidas", los tiempos hasta el evento se establecen, al menos, desde su ingreso hasta el último contacto de observación. Para las unidades que aún están libres del evento, los tiempos al evento se establecen, al menos, desde la entrada hasta el final del estudio. Estas dos últimas observaciones son censuradas. Como los tiempos de entrada no son simultáneos, los tiempos censurados también son diferentes; este es el esquema de censura Tipo III. En resumen, en este esquema de censura los tiempos de censura C_i no son prefijados, sino que son respuestas de variables aleatorias.

Note que este esquema de censura aleatorio incluye el caso de censura Tipo I, donde el tiempo de censura de cada unidad se fija previamente. También incluye el caso donde las unidades, aunque entran al estudio aleatoriamente en el tiempo, se observan hasta un tiempo establecido con anterioridad.

No obstante, se puede advertir que el esquema de censura aleatorio no es suficientemente general, pues no incluye algunos esquemas de censura.

Un ejemplo, tomado de P. J. Smith [64], se refiere a un estudio de 3550 mujeres con edad superior a 65 años, a quienes se les había diagnosticado cáncer de ovario entre los años 1984 y 1989. Los tiempos de sobrevida fueron censurados en las pacientes que estaban vivas al finalizar el año 1990. Los tiempos de censura fueron medidos desde la fecha del diagnóstico de la enfermedad, aunque estas fechas varían considerablemente dado que ellas ingresaron al estudio en varios estadios o etapas de la enfermedad durante el periodo de 6 años (1984-1989). Este es un ejemplo de censura aleatoria a la derecha. La duración limitada del estudio fue considerada como la única causa de censura, hecho que hace razonable suponer que la censura y los tiempos al evento (muerte) son independientes.

1.3.1. Truncamiento

El truncamiento se presenta cuando solo se consideran aquellas unidades cuyos tiempos hasta el evento estén dentro de cierta ventana, $[T_1, T_2]$, de observación. El tiempo para el evento de una unidad que no esté dentro de este intervalo es descartado con toda su información para el análisis; de otra manera, solo se consideran tiempos tales que $T_1 \leq \mathbf{T} \leq T_2$. Esta es una diferencia con la censura, donde se considera para el análisis estadístico una información parcial sobre cada unidad censurada.

Las unidades cuyos tiempos para el evento T exceden al tiempo T_i se dice que tienen *truncamiento a la izquierda*; es decir, $\mathbf{T} \geq T_i$. Una manera de entender este concepto es compararlo con la percepción visual de un objeto, pues solo vemos aquellos cuyo tamaño es mayor o igual a cierto umbral (lo mismo se podría decir para la audibilidad).

El *truncamiento a la derecha* se presenta cuando solo se consideran los tiempos para el evento si este es menor o igual que T_2, es decir, $\mathbf{T} \leq T_d$. Nuevamente, acudiendo a la comparación con el caso de la percepción visual de un objeto, se tiene que este es visible si está a una distancia menor que cierto umbral. Así, se puede afirmar que la visibilidad humana está truncada a la derecha, pues vemos objetos que se encuentren máximo a determinada distancia de nosotros.

Se adopta la siguiente notación para indicar censura a derecha, a izquierda o por intervalo. Para el truncamiento no es necesaria notación alguna, ya que esto se establece en el protocolo del estudio.

- $\mathbf{T}^{+}$: tiempo censurado a derecha.
- $^{-}\mathbf{T}$: tiempo censurado a izquierda.
- $(C_i,\ C_d]$: tiempo censurado por intervalo.
- $\mathbf{T}^{\mathbf{a}}$: abandono o retiro (en inglés, *drop out*) de la unidad.

Tabla 1.1. Tiempo hasta aprender a escribir

Estudiante	Sexo	Tiempo (semanas)	Edad (años)
1	F	16	6
2	F	20	7
3	M	18^{a}	5
4	F	(25, 30]	6
5	M	18	7
6	M	40^{+}	6
7	F	$^{-}12$	7
8	M	33	6
9	M	15^{a}	7
10	F	27	8
11	M	40^{+}	8
12	M	40	8

La tabla 1.1 contiene los datos relacionados con el tiempo para que estudiantes de una institución educativa de carácter público, con similar condición socioeconómica y prácticas pedagógicas, aprendan a escribir. El estudio se inicia desde la semana académica 12 y termina en la semana académica 40; es decir, el estudio se desarrolla dentro del intervalo de tiempo académico [12, 40].

Desde estos datos se observa que los niños 1, 2, 5, 8 10 y 12 muestran la destreza de escritura en un tiempo observado por el docente. Se encuentran datos censurados a derecha para los estudiantes 6 y 11, pues se terminó el periodo de estudio y estos estudiantes no mostraron saber escribir. Los estudiantes 3 y 9 abandonaron sus estudios en la institución por cambio de residencia de sus padres. El estudiante 7 ya sabía escribir cuando se inició el estudio, por ello se considera que tiene censura a izquierda. El estudiante 4

mostró la habilidad y destreza de escribir en la semana 30 pero en su evaluación previa realizada en la semana 25 no había mostrado esta destreza, así que se considera censurado en el intervalo $(25, 30]$.

La teoría del análisis de sobrevida o de confiabilidad se refiere a problemas probabilísticos y estadísticos que están relacionados con las distribuciones de vida de los artículos sujetos a un evento. Los modelos probabilísticos describen el desempeño, el rendimiento y la degradación de las unidades mediante variables aleatorias asociadas con el tiempo al evento. Las distribuciones de probabilidad de estas variables se usan para determinar la función de sobrevida (o confiabilidad), la tasa de falla (o riesgo) y vida media de estos artículos. Los aspectos estadísticos de esta teoría resuelven aspectos de estimación e inferencia de los parámetros de distribución del tiempo al evento.

1.4. Función de sobrevida

La función de sobrevida evalúa la probabilidad de que, en una unidad, el evento ocurra después de un cierto tiempo t o, de manera equivalente, que el evento no se haya presentado sobre la unidad antes del tiempo t; es decir, que permanece en el estado $\mathbf{T}^c$ (figura 1.4). Así, la función de sobrevida o función de confiabilidad, $S(t)$, es definida para los valores de $t \geq 0$ tal que

$$S(t) = P(T > t) = \int_t^{\infty} f(u)du = 1 - F(t).$$

Del teorema fundamental del cálculo, $f(t)dt = -dS(t)$, así, la función de densidad es $f(t) = -\frac{d}{dt}S(t)$.

De la definición de función de sobrevida se desprenden las siguientes propiedades:

i) $S(0) = 1$.

ii) $S(+\infty) = \lim_{t\to+\infty} S(t) = \lim_{t\to\infty} \int_t^{\infty} f(u)du = 0$.

iii) Si $a \leq b$, entonces $S(a) - S(b) = \int_a^b f(u)du \geq 0$.

iv) La función de sobrevivencia S es monótona no creciente.

La tasa de decrecimiento varía de acuerdo con el riesgo que la unidad muestre a la ocurrencia del evento en el tiempo t, pero es difícil determinar, en esencia, el modelo probabilístico asociado con el tiempo al evento solo observando la curva de sobrevivencia. No obstante, el uso de esta curva

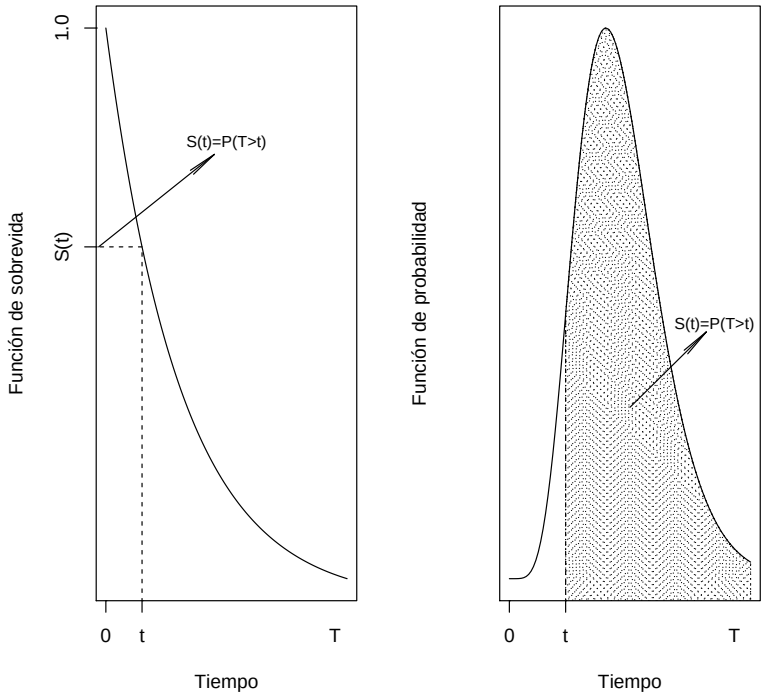

Figura 1.5. Función de sobrevida o de confiabilidad.

aporta, dentro de un análisis exploratorio y descriptivo, elementos para hipotetizar, por lo menos empírica y cualitativamente, sobre la probabilidad de que el evento ocurra después de cierto tiempo.

En una fase exploratoria y descriptiva, es usual comparar dos o más curvas de sobrevivencia para observar, a través del tiempo, el comportamiento diferencial entre ellas, y en consecuencia, buscar las causas atribuibles a las diferencias. No obstante, por las propiedades de $S(t)$ descritas anteriormente, las diferencias no son uniformes ni se mantienen, puesto que las gráficas asociadas con diferentes funciones de sobrevida pueden "cruzarse" o traslaparse. En la figura 1.7 se muestran dos escenarios de funciones de sobrevida, es decir, el "paralelismo" y el "traslape" de las curvas correspondientes a sus funciones de sobrevida. En caso de asumir la variable aleatoria *tiempo*, $\mathbf{T}$, como discreta, esta podría tomar los valores t_j, con $j = 1, 2, \ldots$ y con función de probabilidad $p(t_j) = P(\mathbf{T} = t_j)$, para $j = 1, 2, \ldots$, donde se asume que $t_1 < t_2 < \cdots$. La función de sobrevida para la variable aleatoria discreta

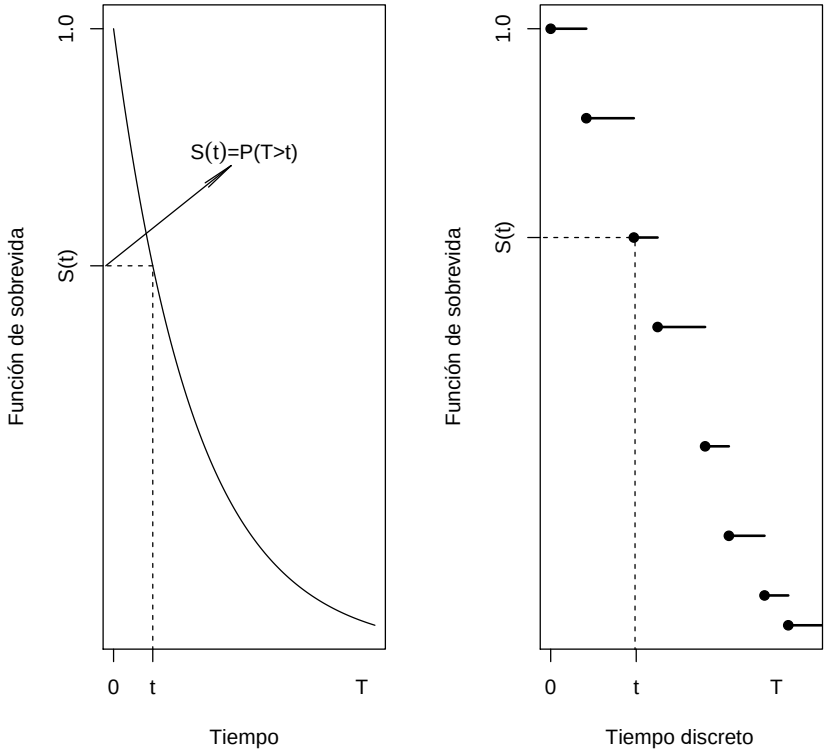

Figura 1.6. Funciones de sobrevida continua y discreta.

T es dada por

$$S(t) = P(T > t) = \sum_{t_j > t} p(t_j).$$

En análisis de sobrevivencia, estas variables surgen cuando los tiempos de falla se agrupan en intervalos, o cuando, por razones logísticas en el proceso de medición, los tiempos al evento hacen referencia a unidades en números enteros no negativos.

Así, aunque se considere la variable aleatoria **T** como continua, toda muestra empírica de datos se puede considerar discreta, pues, en general, es finita. Dadas las variables aleatorias $T_1, \ldots, T_n$ independientes e idénticamente distribuidas, la función de sobrevida empírica es

$$\tilde{S}_n(t) = \frac{No.obs > t}{n} = \frac{1}{n} \sum_{i=1}^{n} I_{[t,\infty)}(t_i), \tag{1.4.1}$$

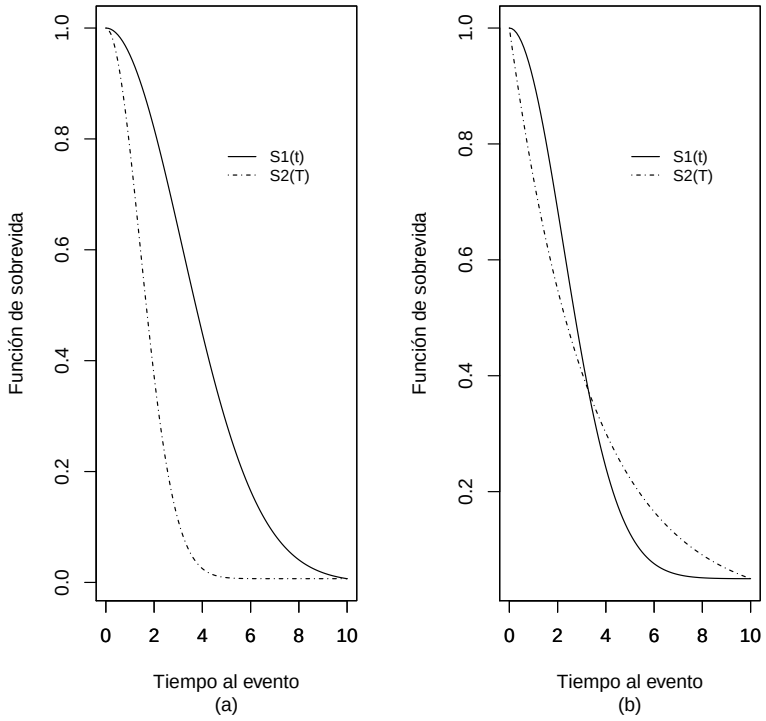

Figura 1.7. Funciones de sobrevida.

la cual es un estimador de $S(t)$.

Desde la expresión (1.4.1), cada sumando $I_{[y,\infty)}(t_i) = Z_i$ toma el valor 1 o 0, según se cumpla la condición de que t_i pertenezca o no al intervalo $[t, \infty)$. Luego,

$$n\tilde{S}_n(t) = \sum_{i=1}^{n} Z_i,$$

es una suma de variables aleatorias con la misma distribución de probabilidad Bernoulli (pues las T_i son independientes e idénticamente distribuidas), con

$$P(Z_i = 1) = P(T_i > t) = S(t) \text{ y } P(Z_i = 0) = P(T_i \leq t) = 1 - S(t).$$

En resumen, la distribución de probabilidad de Z_i, para $i = 1, 2, \ldots, n$ es como se muestra en la siguiente tabla:

z	1	0
$P(Z = z)$	$S(t)$	$1 - S(t)$

Por tener cada Z_i una distribución Bernoulli, la media es $E(Z_i) = S(t)$ y la varianza $var(Z_i) = S(t)[1 - S(t)]$, lo cual significa que $n\tilde{S}_n(t)$ tiene distribución binomial, es decir, $n\tilde{S}_n(t) \sim B(n, S(t))$. En particular, para un valor fijo de t

$$E(\tilde{S}_n(t)) = \frac{1}{n}[nS(t)] = S(t), \tag{1.4.2}$$

$$var(\tilde{S}_n(t)) = \frac{1}{n^2}\{nS(t)[1 - S(t)]\} = \frac{S(t)[1 - S(t)]}{n}. \tag{1.4.3}$$

Un estimador de esta varianza es

$$\widehat{var}(\tilde{S}_n(t)) = \frac{\tilde{S}_n(t)[1 - \tilde{S}_n(t)]}{n}.$$

Para un punto fijo $t = t^*$, un estimador de la probabilidad de que el evento ocurra después del tiempo t^* se obtiene mediante $\tilde{S}_n(t^*)$. Como se muestra en la expresión (1.4.2), $\tilde{S}_n(t^*)$ es un estimador insesgado de $S(t^*)$. La precisión de esta estimación se mide mediante un intervalo del 95 % de confianza con la expresión

$$\tilde{S}_n(t^*) \mp 2\sqrt{\frac{\tilde{S}_n(t^*)[1 - \tilde{S}_n(t^*)]}{n}}.$$

La comparación de las funciones de sobrevida no siempre es inmediata y concluyente desde un punto de vista gráfico, pues todo gráfico es visualmente finito (en un rango de valores de $\mathbf{T}$), por lo que habrán características que un gráfico no revela al observador. Es decir, puede pasar que dos funciones de sobrevida S_1 y S_2 sean tales que, para algún tiempo t^*, en un intervalo de tiempo (t_1, t_2), la función de sobrevida $S_1(t^*)$ esté por encima de la función de sobrevida $S_2(t^*)$, es decir, $S_1(t^*) > S_2(t^*)$. Esta desigualdad puede ser la recíproca en puntos del tiempo fuera de este intervalo como se muestra en la figura 1.7.

Una alternativa para comparar curvas de sobrevida es hacerlo a través de las medias de la variable aleatoria *tiempo para el evento*, $\mathbf{T}$.

Para la variable aleatoria $\mathbf{T}$ su media (valor esperado), μ, o tiempo medio para el evento o falla (TMF), es

$$E(\mathbf{T}) = \mu = \int_0^\infty tf(t)dt.$$

Mediante el método de integración por partes se tiene que

$$E(\mathbf{T}) = \mu = \int_0^\infty S(t)dt.$$

La cual se puede interpretar geométricamente como el área bajo la curva $S(t)$ en todo el soporte de esta función.

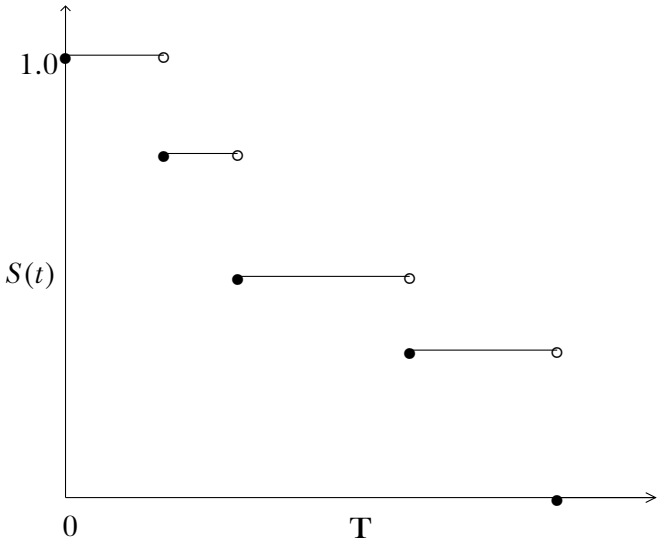

Figura 1.8. Función de sobrevida discreta.

1.5. Función de riesgo

La función de riesgo se define como la probabilidad de falla durante un intervalo de tiempo muy pequeño, suponiendo que sobre la unidad de estudio el evento no ha ocurrido hasta el inicio del intervalo. También se entiende como el límite de la probabilidad de que a una unidad le ocurra el evento en un intervalo muy corto, t a $t+\delta t$, dado que la unidad no ha tenido el evento hasta el tiempo t.

Así, la *v. a.* $\mathbf{T}$ tiene *función de riesgo (riesgo instantáneo)*, $h(t)$,[1] para $t > 0$,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq \mathbf{T} \geq t + \Delta t | \mathbf{T} > t)}{\Delta t} = \frac{\lim\limits_{\Delta t \to 0} \frac{P(t \leq \mathbf{T} \leq t + \Delta t)}{\Delta t}}{P(\mathbf{T} > t)}$$
$$= \frac{\lim\limits_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t}}{P(T > t)} = \frac{\frac{d(F(t))}{dt}}{S(t)} = \frac{f(t)}{S(t)}. \qquad (1.5.1)$$

Esto indica la disposición inmediata al evento en un intervalo de tiempo pequeño, como una función de la edad, t, sobre una unidad u objeto.

Una función de riesgo creciente advierte que la tasa del evento de una unidad aumenta con el tiempo, en consecuencia, la unidad es cada vez más vulnerable a que experimente el evento. Esto ocurre sobre algunas unidades

[1] Se conserva la inicial de la palabra en inglés *hazard*, que significa *riesgo*.

debido al debilitamiento por envejecimiento (vetustez o fatiga). Una función de riesgo constante advierte que la tasa de riesgo no cambia con el paso del tiempo. Una función decreciente muestra que el riesgo al evento disminuye a medida que el tiempo pasa. En la figura 1.9 se muestran estos tres tipos de escenarios de riesgo mediante las respectivas funciones, $h(t)$, de riesgo.

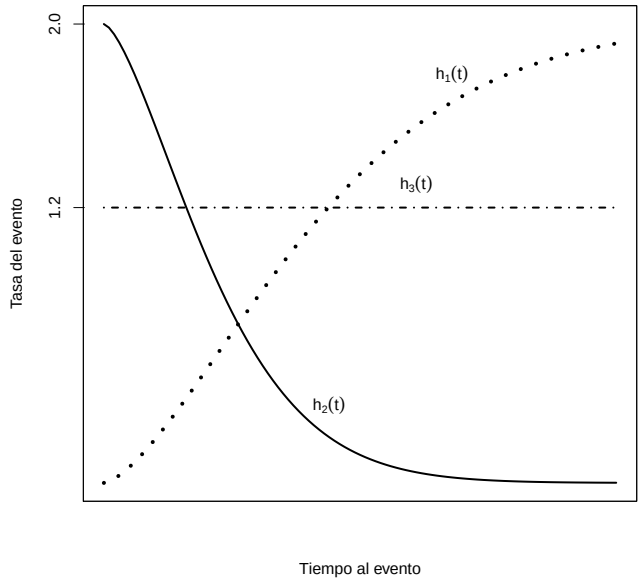

Figura 1.9. Funciones de riesgo.

Cuando el objeto o propósito del estudio es la evolución de un ser vivo, se puede estimar esta función considerando una serie de individuos en condiciones similares, de modo que, si se quiere modelar el tiempo de vida, una estimación natural de la función de riesgo es el número de eventos ocurridos por unidad de tiempo en cada intervalo de observación. Con el propósito de tener un modelo adecuado mediante la función de riesgo, $h(t)$, se debe considerar la existencia de tres formas o circunstancias de ocurrencia del evento, las cuales están asociadas a características esencialmente temporales.

- El primer tipo, llamado *evento inicial*, se puede presentar al principio de la vida del individuo y va desapareciendo conforme se desarrolla el periodo inicial. Un ejemplo de este tipo puede observarse en las tablas de mortalidad humana, según las cuales se supone que, en la vida de un individuo, hay presentes, al principio, ciertos problemas de

carácter hereditario que pueden provocar desenlaces fatales, que van desapareciendo conforme el individuo crece y que están presentes, aproximadamente, hasta una edad de 10 años. Una situación similar se da en la mayoría de animales y plantas.

- El segundo tipo se llama *evento accidental*, el cual puede ocurrir durante el periodo en el que el individuo presenta una función de riesgo constante, generalmente menor que la prevaleciente durante su periodo inicial. En las tablas de mortalidad humana, se supone que las muertes ocurridas entre los 10 y 30 años están asociadas, principalmente, con situaciones accidentales.
- El tercer tipo, llamado *evento de desgaste*, se asocia con un deterioro gradual del individuo (fatiga). En las tablas de mortalidad humana hay, a partir de los 30 años, una proporción creciente de muertes atribuidas al envejecimiento del individuo. Si en el proceso de desarrollo de un individuo únicamente estuviesen presentes estos tres tipos de circunstancias para el evento, el modelo seleccionado para representar el tiempo de funcionamiento debe tener una función de riesgo que, por su forma, es conocida con el nombre de curva de "bañera", y se asemeja a la mostrada en la figura 1.10. Este tipo de función reúne, simultáneamente, las tres características de función de riesgo (decreciente, constante y creciente).

En consecuencia, desde (1.5.1), se tiene la siguiente relación entre la función de probabilidad, la función de sobrevida y la función de riesgo:

$$h(t) = \frac{f(t)}{S(t)}. \tag{1.5.2}$$

Además, como

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} = \frac{-d(\ln S(t))}{dt}, \tag{1.5.3}$$

entonces se tiene la siguiente expresión para la función de sobrevida:

$$S(t) = e^{-\int_0^t h(u)du} = e^{-H(t)}, \tag{1.5.4}$$

con

$$H(t) = \int_o^t h(u)du, \tag{1.5.5}$$

la igualdad (1.5.4) equivale a

$$H(t) = -\ln(S(t)), \tag{1.5.6}$$

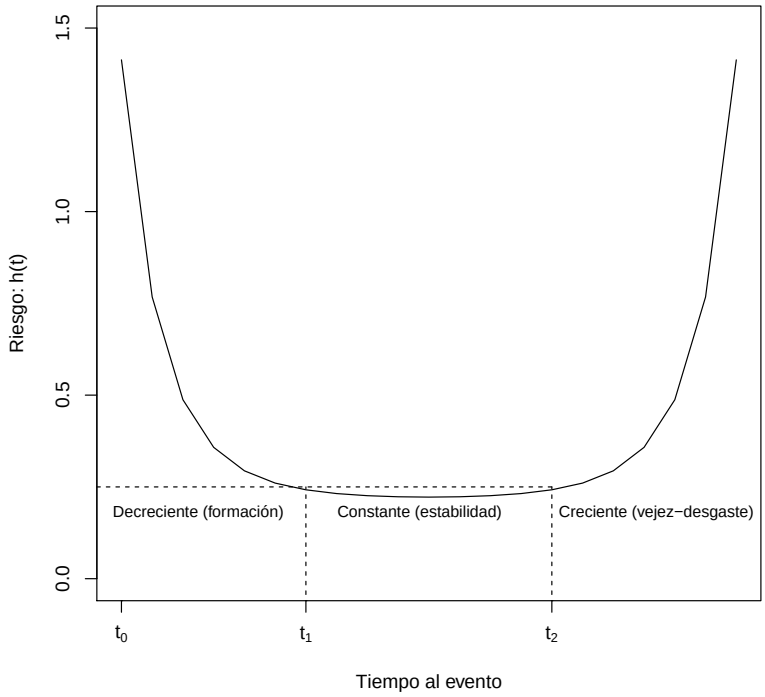

Figura 1.10. Función de riesgo tipo "bañera".

las cuales son formas equivalentes a *la función de riesgo acumulado*, para **T** variable aleatoria continua. Para el caso discreto, la función de riesgo acumulado es

$$H(t) = \sum_{t_j < t} h(t_j). \tag{1.5.7}$$

La función de riesgo es también conocida como *la tasa instantánea de mortalidad*, la fuerza de mortalidad, la tasa de mortalidad condicional, la tasa de falla de edad específica y demás nombres relacionados con el tema que se esté tratando, además de la interpretación que se tenga dentro de este. Es una medida de propensión al evento en función de la edad de la unidad, en el sentido de que la cantidad $\Delta t h(t)$ es la proporción esperada de individuos de edad t que les ocurrirá el evento en el intervalo t a $t + \Delta t$. También puede verse como una aproximación a la probabilidad de que un individuo de edad t experimente el evento en el siguiente instante. En algunos casos, como en el área de la confiabilidad, para ingeniería, se interpreta también

como la velocidad de degradación de un dispositivo, o la intensidad con que se presenta el evento en el dispositivo justo en el instante t. Por eso la estimación de la función de riesgo se basa esencialmente en consideraciones físicas.

Ejemplo 1.5.1. La distribución exponencial (se considera en el capítulo 3) describe adecuadamente el patrón de *tiempo hasta el evento* en dispositivos o sistemas electrónicos o mecánicos. El papel de la distribución exponencial es similar al que tiene la distribución normal en varias áreas de la estadística.

Cuando la variable *tiempo* $\mathbf{T}$ sigue la distribución exponencial con parámetro θ, se escribe $\mathbf{T} \sim \exp(\theta)$. La función de densidad de probabilidad es definida por

$$f(t,\theta) = \begin{cases} \theta e^{-\theta t}, & t > 0, \theta > 0 \\ 0, & t < 0. \end{cases}$$

La función de distribución acumulada es

$$F(t) = 1 - e^{-\theta t},\ t \geq 0.$$

La función de sobrevida es

$$S(t) = e^{-\theta t}.$$

La función de riesgo, de acuerdo con (1.5.2), es

$$h(t) = \frac{f(t)}{S(t)} = \frac{\theta e^{-\theta t}}{e^{-\theta t}} = \theta, t \geq 0.$$

Esta función indica que el riesgo es constante e independiente del tiempo, lo cual permite afirmar, por ejemplo para el caso de unidades eléctricas, que no hay fatiga, desgaste o degradación por su uso. Esta situación no ocurre en humanos ni, en general, en los seres vivos. La media de una variable aleatoria $\mathbf{T}$ con distribución exponencial de parámetro θ es $E(\mathbf{T}) = \theta^{-\mathbf{1}}$ y su varianza es $var(\mathbf{T}) = \theta^{-\mathbf{2}}$.

En el caso discreto, sea T una $v.\ a.$ discreta que toma valores t_j con $j = 1, 2, \ldots$ La función de riesgo se define para los valores t_j y proporciona la probabilidad condicional de ocurrencia del evento al tiempo $\mathbf{T} = t_j$; dado que a la unidad no le ha ocurrido el evento antes de t_j, por lo tanto, se tiene que

$$h(t_j) = P(\mathbf{T} = t_j \mid \mathbf{T} \geq t_j) \tag{1.5.8}$$

$$= \frac{P(\mathbf{T} = t_j)}{P(\mathbf{T} \geq tj)} \tag{1.5.9}$$

$$= \frac{P(tj)}{S(t_{j^-})}, \tag{1.5.10}$$

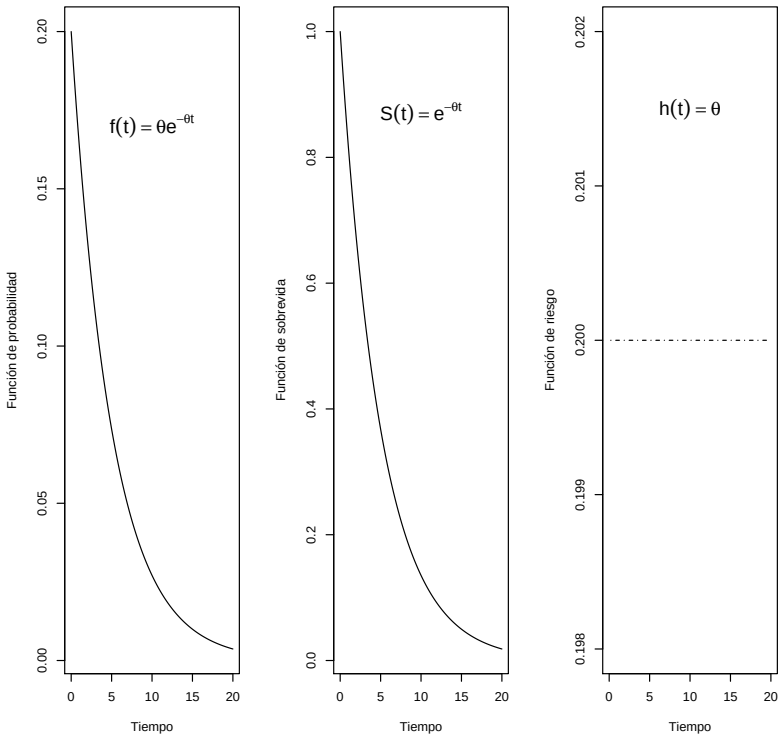

Figura 1.11. Función de probabilidad, sobrevida y riesgo en exp(θ).

donde t_{j^-} corresponde a un instante inmediatamente antes de t_j y, por tanto, $P(\mathbf{T} \geq t_j) = 1 - P(\mathbf{T} < t_j) = S(t_{j^-}) \neq S(t_j)$ en el caso discreto.

La función de riesgo describe la forma como cambia la tasa instantánea de la ocurrencia de un evento de interés en el transcurso del tiempo. La única restricción para esta función es que tiene que ser no negativa, es decir, $h(t) \geq 0$. La función de riesgo puede crecer, decrecer, permanecer constante o tener un proceso más complejo. En el capítulo 3 se muestran las gráficas para varias funciones de probabilidad en valores distintos de los parámetros que las identifican.

En ambos casos, tanto en el discreto como en el continuo, la función, $H(t)$, como su nombre lo indica, acumula el riesgo con el paso del tiempo. De tal manera, corresponde a una función no decreciente, y de acuerdo con su forma de incrementarse, se podría tener información del comportamiento del riesgo a lo largo del tiempo.

El *valor esperado* y *la varianza* de la variable aleatoria *tiempo al evento*, **T**, son, respectivamente,

$$\mu = E(\mathbf{T}) = \int_0^\infty uf(u)du = \int_0^\infty S(u)du \text{ y} \tag{1.5.11}$$

$$\sigma^2 = var(\mathbf{T}) = 2\int_0^\infty uS(u)du - \left[\int_0^\infty S(u)du\right]^2. \tag{1.5.12}$$

El interesado en la obtención de las expresiones (1.5.11) y (1.5.12) puede hacerlo mediante integración por partes.

El p-ésimo *cuantil*, ξ_p, con $0 < p < 1$ (también referido como el percentil 100p-ésimo), de la variable aleatoria **T** es $\xi_p = F^{-1}(p)$. También se define como el valor más pequeño ξ_p tal que

$$S(\xi_p) \leq 1 - p, \text{ es decir, } \xi_p = \text{ínf}\{t : S(t) \leq 1 - p\}. \tag{1.5.13}$$

Si la variable aleatoria **T** es continua, el p-ésimo cuantil se encuentra como solución de la ecuación $S(\xi_p) = 1 - p$. Así, la mediana de la distribución de la variable aleatoria *tiempo para el evento* es $\xi_{0.5}$ tal que $S(\xi_{0.5}) = 0.5$, y el tercer cuartil, $\xi_{0.75}$, es tal que $S(\xi_{0.75}) = 0.75$, entre otros.

Ejemplo 1.5.2. Para una variable **T** con distribución exponencial de parámetro θ ($\mathbf{T} \sim \exp(\theta)$), la media y la mediana de acuerdo con (1.5.11) y (1.5.13) son

$$\mu = \int_0^\infty S(u)du = \int_0^\infty e^{-\theta u}du = \frac{1}{\theta}$$

$$S(\xi_{0.5}) = \frac{1}{2},$$

$$e^{-\theta\xi_{0.5}} = \frac{1}{2}, \text{ luego } \xi_{0.5} = \ln 2/\theta.$$

La *función de vida media residual* al tiempo t_0, también llamada la *media de vida residual*, **MVR**, para las unidades con edad t_0, mide la esperanza de tiempo de vida restante ($\mathbf{T} - t_0$) o el tiempo esperado antes de la ocurrencia del evento de interés.

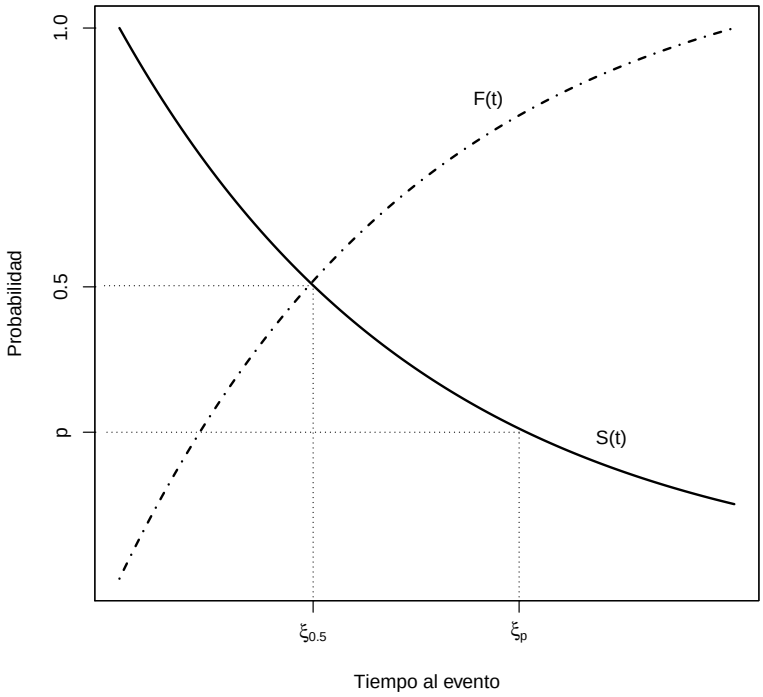

Figura 1.12. Cuantiles.

Para $t_0 > 0$, la función de vida media residual a la edad t_0, $r(t_0)$, es

$$\begin{aligned} MVR = r(t_0) &= E(\mathbf{T} - t_0 \mid \mathbf{T} > t_0) \\ &= \int_{t_0}^{\infty} (t - t_0) dP(t_0 \leq \mathbf{T} \leq t \mid \mathbf{T} > t_0) \\ &= \int_{t_0}^{\infty} (t - t_0) d\left(\frac{S(t_0) - S(t)}{S(t_0)}\right) \\ &= \int_{t_0}^{\infty} (t - t_0) \left(\frac{-S'(t)dt}{S(t_0)}\right) \\ &= \frac{-(t - t_0)S(t)|_{t_0}^{\infty} + \int_{t_0}^{\infty} S(t)dt}{S(t_0)} \\ &= \frac{\int_{t_0}^{\infty} S(t)dt}{S(t_0)}. \end{aligned}$$

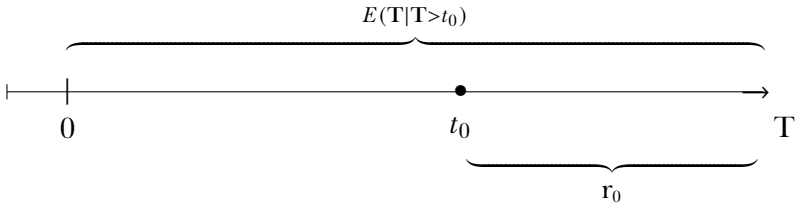

Figura 1.13. Media residual y esperanza de vida a la edad t_0.

Así,

$$r(t_0) = \frac{\int_{t_0}^{\infty} S(t)dt}{S(t_0)}. \tag{1.5.14}$$

Se demuestra que la **MVR** al tiempo inicial $r(0) = E(T)$.

El cálculo de la función de vida media residual empírica, de acuerdo con la expresión (1.4.1), es

$$\tilde{r}(t_0) = \frac{\sum\limits_{t>t_0} \tilde{S}_n(t)}{\tilde{S}_n(t_0)}, \tag{1.5.15}$$

donde

$$\tilde{S}_n(t) = \frac{No.obs > t}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{[y,\infty)}(t_i).$$

La figura 1.13 muestra la media residual y la esperanza de vida a la edad t_0.

Para una *v. a.* de sobrevida **T** y un tiempo fijo t_0, *la esperanza de vida a la edad* t_0 es una medida del lapso de tiempo al evento, dado que se ha alcanzado a la edad t_0. Esta es dada por

$$E(\mathbf{T} \mid \mathbf{T} > t_0) = t_0 + r(t_0) = t_0 + \frac{\int_{t_0}^{\infty} S(t)dt}{S(t_0)}. \tag{1.5.16}$$

Ejemplo 1.5.3. La media residual de vida esperada a la edad t_0 se presenta en la tabla 1.2 para hombres y bombillos.

Los datos advierten que en los humanos, representados por los hombres, la esperanza de sobrevida futura (residual) a la edad t_0 se reduce en la medida en que aumenta la edad t_0, mientras que para los bombillos esta se mantiene constante, pues, para estos, se asume que la variable aleatoria *tiempo para la muerte* tiene una distribución exp(73.3), de manera que el cálculo se hace mediante la expresión (1.5.14):

$$r(t_0) = \frac{\int_{t_0}^{\infty} S(t)dt}{S(t_0)} = \frac{\int_{t_0}^{\infty} e^{-u/73.3}du}{e^{-t_0/73.3}} = \frac{73.3e^{-t_0/73.3}}{e^{-t_0/73.3}} = 73.3,$$

Tabla 1.2. Esperanza de vida residual en t

Edad actual t_0	Hombre (años)	Bombillo (Unidades de tiempo)
0	73.3	73.3
10	64.2	73.3
20	54.7	73.3
30	45.2	73.3
40	35.9	73.3
50	26.7	73.3
60	18.3	73.3
70	11.5	73.3
80	6.5	73.3
90	3.5	73.3

Tomado de Smith [64, p. 12].

esto señala que la esperanza de vida residual es independiente de t_0, es decir, es constante e igual a 73.3 unidades de tiempo como se muestra en la tabla 1.2 (última columna).

Los valores de la esperanza de vida residual en humanos se obtienen de los cálculos en una tabla de mortalidad (calculados mediante (1.4.1)). En particular, al nacer, un hombre tiene una esperanza de vida residual de 73.3 años y un bombillo nuevo (no usado) tiene una esperanza de vida residual de 73.3 unidades de tiempo. A los 90 años, un hombre tiene una esperanza de vida residual de 3.5 años (es decir, lo que se espera que "le faltaría por vivir"), mientras que una bombilla, después de 90 unidades de tiempo en uso, se espera que funcione 73.3 unidades de tiempo antes de fallar.

Tabla 1.3. Vida futura esperada a la edad t_0 en humanos versus objeto

Edad actual	Hombre	Mujer	Bombillo
0	73.3	79.5	73.3
10	74.2	80.3	83.3
20	74.7	80.5	93.3
30	75.2	80.7	103.3
40	75.9	81.0	113.3
50	76.7	81.6	123.3
60	78.3	82.8	133.3
70	81.5	84.8	143.3
80	86.5	88.4	153.3
90	93.5	94.1	163.3

Tomado de Smith [64, p. 11].

En la tabla 1.3 se muestran los datos para las esperanzas de vida en humanos (hombres y mujeres) frente a la esperanza de vida de un tipo de

bombillo. Por la notación numérica los tiempos parecen los mismos para la edad de las personas y la vida útil de las bombillas, pero son en esencia tiempos diferentes, ya que en el caso de las personas se miden en tiempo calendario, mientras que en el caso de las bombillas se miden en unidades de uso, que pueden ser unidades de tiempo (por ejemplo, 10 o 100 horas de funcionamiento, según el tipo de bombilla).

Capítulo

dos

Métodos no paramétricos para estimar la función de sobrevida

$q(j) = P[t(j) \in [t(j-1)\ \ q(j) = P[t(j) \in [t(j-1)\ \ q(j) = P[t(j) \in [t(j-1)\ \ q(j) = P[t(j) \in [t(j-1)\ \ q(j) = P[t(j) \in [t(j-1)\ \ q$

2.1. Introducción

En el análisis de datos de tiempo hasta un evento, se presentan estadísticas que resumen la información, como la media, la mediana, el error estándar para la media, los percentiles y cuantiles, entre otras. Sin embargo, debido a la posible presencia de censuras, las estadísticas que resumen la información pueden no tener las propiedades deseadas, tales como el insesgamiento, la varianza mínima y la consistencia. Por ejemplo, la media muestral para estos datos deja de ser un estimador insesgado de la media poblacional del tiempo hasta el evento. De esta forma, es necesario usar otros métodos para explorar, describir y estimar la distribución de probabilidad, la función de sobrevida y la función de riesgo subyacente sobre la base de un conjunto de datos con censura o truncamiento. Cuando esta distribución es estimada, se pueden, entonces, estimar otras cantidades de interés tales como la media, la mediana, etc.

Los métodos de estimación que se desarrollan en este capítulo no requieren suponer un modelo probabilístico para la variable *tiempo hasta el evento*, de aquí el calificativo de no "paramétricos".

2.2. Caso sin censura

Suponga que se tiene una muestra de tiempos, $T_1, T_2, \ldots, T_n$, hasta un evento donde ninguna de las observaciones está censurada. Los tiempos ordenados son $T_{(1)} \leq T_{(2)} \leq, \ldots, \leq T_{(n)}$; con $T_{(1)} = min\{T_1, T_2, \ldots, T_n\}$ y $T_{(n)} = max\{T_1, T_2, \ldots, T_n\}$. La función de sobrevivencia, $S(t)$, es la probabilidad de que a una unidad le ocurra el evento en un tiempo mayor que t. Esta función puede ser estimada por la función de sobrevivencia empírica dada por (1.4.1),

$$\tilde{S}_n(t) = \frac{No.obs > t}{n} = \frac{\text{Núm. de unidades que sobreviven más allá de } t}{\text{Núm. total de unidades en el conjunto de datos}}$$
$$= \frac{1}{n}\sum_{i=1}^{n} I_{[t,\infty)}(T_i).$$

De manera equivalente, $\tilde{S}_n(t) = 1 - \tilde{F}_n(t)$, donde $\tilde{F}_n(t)$ es la función de distribución empírica, la cual es el número de unidades sobre las cuales el evento ha ocurrido al tiempo t dividido entre el número total de unidades. De esta definición empírica para la función de sobrevida, se tiene que $\tilde{S}(t) = 1$ para valores de t menores que el mínimo $T_{(1)}$ (antes que el evento ocurra en alguna unidad) y que $\tilde{S}(t) = 0$ después del tiempo del último

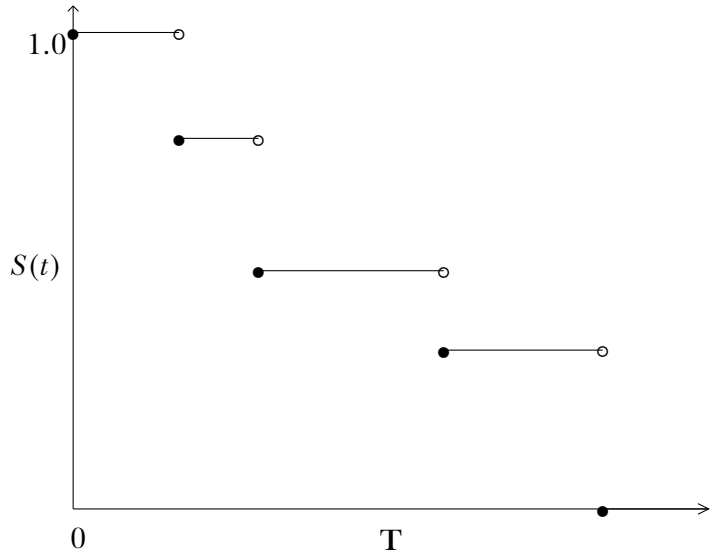

Figura 2.1. Función de sobrevida discreta.

evento; es decir, mayores o iguales que el máximo $T_{(n)}$. La función de sobrevida estimada $\tilde{S}(t)$ es considerada constante entre dos tiempos de falla adyacentes, por lo que la gráfica de $\tilde{S}(t)$ es una función escalonada que decrece inmediatamente después de cada tiempo al evento observado.

Una gráfica de la función de sobrevida se muestra en la figura 2.1. Se observa que es una función escalonada no creciente.

2.3. Estimador de Kaplan-Meier

En esta sección se muestra la estructura y el procedimiento para la construcción del estimador de la función de sobrevida mediante el estimador de Kaplan–Meier (en adelante KM); para ello se sigue la construcción desarrollada en el libro de Collet [10]. Para determinar el estimador de Kaplan-Meier (1958) de una función de sobrevivencia a partir de una muestra que contiene datos censurados por la derecha, se construyen intervalos de tiempo adyacentes, cada uno de los cuales contiene un tiempo al evento. Se considera que esta falla ocurre al inicio del intervalo.

Por ejemplo, sean $t_{(1)}$, $t_{(2)}$ y $t_{(3)}$ tres tiempos observados, ordenados de tal forma que $t_{(1)} < t_{(2)} < t_{(3)}$ y sea C un tiempo para el evento censurado que ocurre entre $t_{(2)}$ y $t_{(3)}$. Los intervalos construidos comienzan en $t_{(1)}$, $t_{(2)}$ y $t_{(3)}$, y cada uno incluye el tiempo para un evento, aunque pueden ha-

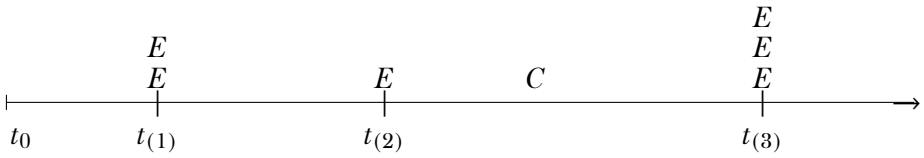

Figura 2.2. Construcción de intervalos en estimador Kaplan–Meier.

ber más de una unidad a la que le ocurra el evento en cualquier tiempo al evento en particular. En el tiempo de censura, C, ningún intervalo comienza.

Esta situación es ilustrada en la figura 2.2, en la cual **E** representa un evento y C un tiempo censurado. En esta figura se tiene que dos a unidades el evento les ocurre en $t_{(1)}$, a una en $t_{(2)}$ y a tres les ocurre en $t_{(3)}$. El tiempo de origen es denotado por t_0, en el cual comienza el periodo y termina justo antes de $t_{(1)}$, que es el tiempo del primer evento. Esto significa que el intervalo desde t_0 hasta $t_{(1)}$ no incluirá ningún evento. El primer intervalo construido comienza en $t_{(1)}$ y termina justo antes de $t_{(2)}$, este intervalo contiene, por tanto, el tiempo al evento $t_{(1)}$. El segundo intervalo comienza en $t_{(2)}$ y termina justo antes de $t_{(3)}$ e incluye el tiempo al evento $t_{(2)}$ y el tiempo de censura C. El tercer intervalo comienza en $t_{(3)}$ y contiene al máximo, $t_{(3)}$, de los tiempos para el evento.

En general, sean n unidades sobre las cuales son observados los tiempos para el evento $t_1, t_2, \ldots, t_n$. Algunas de estas observaciones pueden estar censuradas por la derecha y también puede suceder que más de una unidad presente el mismo tiempo para el evento (empates). Por tanto, sea r el número de tiempos diferentes entre las unidades, donde $r \leq n$. Los r tiempos al evento son ordenados de forma ascendente $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$; el j-ésimo tiempo hasta el evento se nota por $t_{(j)}$ con $j = 1, 2, \ldots, r$. El número de unidades a las que no les ha ocurrido el evento justo antes del tiempo $t(j)$, incluyendo aquellas que están próximas al evento en este tiempo, se notan por $n_{(j)}$ para $j = 1, 2, \ldots, r$, y e_j nota el número de veces que ha ocurrido el evento al tiempo $t_{(j)}$. El intervalo de tiempo desde $t_{(j)} - \delta$ hasta $t_{(j)}$, donde δ es un lapso de tiempo infinitesimal ($\delta \to 0$), tiene incluido un tiempo al evento. Como hay $n_{(j)}$ unidades que justo antes del tiempo $t_{(j)}$ no les ha ocurrido el evento y se han presentado $e_{(j)}$ eventos al tiempo $t_{(j)}$, la probabilidad de que a una unidad le ocurra el evento en el intervalo $t_{(j)} - \delta$ a $t_{(j)}$ es estimada por $e_{(j)}/n_{(j)}$. El correspondiente estimador de la probabilidad de sobrevivir a través de este intervalo es, entonces, la probabilidad de su complemento, esta es $(n_{(j)} - e_{(j)})/n_{(j)}$.

Algunas veces se presentan observaciones censuradas al mismo tiempo en que ocurre algún evento, de modo que el tiempo al evento y el tiempo censurado ocurren simultáneamente. En este caso, cuando se calcula el valor de $n_{(j)}$, el tiempo censurado es tomado como si ocurriera inmediatamente después del tiempo al evento.

Dado el modo en que los intervalos de tiempo han sido construidos, el intervalo de $t_{(j)}$ a $t_{(j+1)} - \delta$, es decir, el tiempo inmediatamente antes del siguiente tiempo al evento, no contiene eventos. Por lo tanto, la probabilidad de sobrevivencia desde $t_{(j)}$ hasta $t_{(j+1)} - \delta$ es uno, y la probabilidad conjunta de sobrevivir desde $t_{(j)} - \delta$ hasta $t_{(j)}$ y desde $t_{(j)}$ hasta $t_{(j+1)} - \delta$ (intersección de eventos) puede ser estimada por

$$(n_{(j)} - e_{(j)})/n_{(j)}. \tag{2.3.1}$$

En el límite, cuando δ tiende a cero ($\delta \to 0$), $(n_{(j)} - e_{(j)})/n_{(j)}$ tiende a ser un estimador de la probabilidad de sobrevivir desde $t_{(j)}$ hasta $t_{(j+1)}$. Bajo el supuesto de que los eventos sobre las unidades en la muestra ocurren independientemente, el estimador para la función de sobrevida para cualquier tiempo en el k-ésimo intervalo de tiempo construido desde $t_{(k)}$ hasta $t_{(k+1)}$, $k = 1, 2, \ldots, r$, donde $t_{(r+1)}$ es definido como ∞, será la probabilidad estimada de que el evento ocurra más allá de $t_{(k)}$. Esta es la probabilidad de sobrevivir a través del intervalo desde $t_{(k)}$ hasta $t_{(k+1)}$ y, por supuesto, en todos los intervalos anteriores, tal como se hace en Collet [10].

A partir de las consideraciones anteriores, se puede resumir la estimación de la función de sobrevida, $S(t)$, en los siguientes pasos mediante el estimador de **Kaplan-Meier**:

- Se ordenan los tiempos $t_{(1)} < t_{(2)} < \ldots t_{(r)}$ en r tiempos diferentes.
- Se cuenta el número de eventos, $e_{(j)}$, en $t_{(j)}$, con $j = 1, 2, \ldots, r$.
- Se cuenta el número de unidades, $n_{(j)}$, en riesgo al evento en el tiempo $t_{(j)}$, es decir, las unidades que no han experimentado el evento ni son censuradas en el instante inmediatamente anterior a $t_{(j)}$.

En consecuencia, el estimador de **Kaplan-Meier** de la función de sobrevida está dado por

$$\begin{aligned}\tilde{S}_{KM}(t) &= \prod_{j:t_{(j)}<t}\left(\frac{n_{(j)}-e_{(j)}}{n_{(j)}}\right)\\ &= \prod_{j:t_{(j)}<t}\left(1-\frac{e_{(j)}}{n_{(j)}}\right)\\ &= \prod_{j:t_{(j)}<t}(1-q_{(j)}).,\end{aligned} \tag{2.3.2}$$

donde $q_{(j)}$ es la probabilidad que a una unidad le ocurra el evento en el periodo $[t_{(j-1)}; t_{(j)})$, dado que antes del tiempo $t_{(j-1)}$ no le ha ocurrido; es decir,

$$q_{(j)} = P[t_{(j)} \in [t_{(j-1)}; t_{(j)}) | \mathbf{T} \geq t_{(j-1)}],$$

para $t_{(k)} \leq t < t_{(k+1)}$, $k = 1, 2, \ldots, r$, con $\tilde{S}(t) = 1$ para $t < t_{(1)}$ y donde $t_{(r+1)}$ se considera como ∞. En sentido estricto, si la observación más grande es un tiempo de supervivencia censurado, t^*, se dice que $\tilde{S}_{KM}(t)$ está indefinida para $t > t^*$. Por otro lado, si la observación más grande es un tiempo observado, $t_{(r)}$, es una observación no censurada, $n_{(r)} = e_{(r)}$; entonces, $\tilde{S}_{KM}(t)$ es cero para $t > t_{(r)}$. La gráfica del estimador Kaplan-Meier para la función de sobrevivencia es una función escalonada en la cual la probabilidad de sobrevivencia estimada es constante entre tiempos adyacentes y decrecientes en cada tiempo del evento. El estimador Kaplan-Meier también es conocido como el *estimador producto límite*. Es conveniente notar que si no hay tiempos censurados en el conjunto de datos, $n_{(j)} - e_{(j)} = n_{(j+1)}$, $j = 1, 2, \ldots, k$ y por tanto se tiene que

$$\begin{aligned}\tilde{S}_{KM}(t) &= \prod_{j=1}^{k}\left(\frac{n_{(j)}-e_{(j)}}{n_{(j)}}\right)\\ &= \prod_{j=1}^{k}\frac{n_{(j+1)}}{n_{(j)}} = \frac{n_{(2)}}{n_{(1)}} \times \frac{n_{(3)}}{n_{(2)}} \times \cdots \frac{n_{(k+1)}}{n_{(k)}}\\ &= \frac{n_{(k+1)}}{n_{(1)}},\end{aligned}$$

para $k = 1, 2, \ldots, r$, donde $\tilde{S}_{KM}(t) = 1$ para $t < t_{(1)}$ y $\tilde{S}_{KM}(t) = 0$ para $t > t_{(r)}$. Ahora $n_{(1)}$ es el número de unidades expuestas al evento justo antes del primer tiempo para el evento, el cual es el número de unidades en la muestra y $n_{(k+1)}$ es el número de unidades tales que el tiempo para el evento

es mayor o igual a $t_{(k+1)}$. Así, se tiene que en ausencia de censura, $\tilde{S}_{KM}(t)$ es simplemente la función de sobrevida empírica definida anteriormente.

Ejemplo 2.3.1. Los datos corresponden a los tiempos a la muerte de pacientes con hepatitis viral aguda [11]. Los pacientes recibieron de manera aleatoria un placebo o un tratamiento con esteroide (15 y 14 pacientes, respectivamente). A cada paciente se le hizo un seguimiento durante 16 semanas hasta la ocurrencia del evento (muerte) o hasta la pérdida del seguimiento. Los datos están contenidos en la tabla 2.1, el superíndice $^+$ significa censura a derecha. Se consideran los datos del grupo esteroide.

Un procedimiento para obtener el estimador de Kaplan-Meier, como se describió anteriormente, implica una serie de pasos, en los cuales cada paso depende del anterior. Los tiempos observados de sobrevida diferentes (no empatados) son las semanas 1, 5, 7, 8 y 10. Los pasos son generados a partir de los intervalos definidos por el ordenamiento de los tiempos a la muerte, de manera que cada intervalo empieza en un tiempo observado y termina en el siguiente tiempo observado.

A manera de ilustración, para que una persona sobreviva hasta la semana 5, debe sobrevivir a la primera semana y luego a la quinta, dado que sobrevivió a la primera. Así, para calcular la sobrevida a la semana 5 se tiene que:

$$\widetilde{S}(5) = P(\mathbf{T} \geq 5) = P(\mathbf{T} \geq 1, \mathbf{T} \geq 5) = P(\mathbf{T} \geq 1)P(\mathbf{T} \geq 5|\mathbf{T} \geq 1).$$

La tabla 2.2 muestra los tiempos ordenados, $t_{(j)}$, con sus respectivos intervalos, $[t_{(j)}, t_{(j+1)})$, número de muertos, e_j, número de sobrevivientes, $n_{(j)}$, y el valor de la sobrevida, $\widetilde{S}_{KM}(t)$. Los intervalos inician en $[0, 1)$ y terminan en $[10, 16)$; el límite superior del último intervalo es 16, pues este es el mayor tiempo de acompañamiento u observación de un paciente en el estudio.

Al momento $t = 0$ todos los individuos estaban vivos. Así, una estimación de la sobrevida en el intervalo $[0, 1)$ es 1, es decir, $S(t) = 1$ para $t \in [0, 1)$. Al término de 1 semana o $t = 1$, el valor de la sobrevida decae debido a la ocurrencia de tres muertes. En el segundo intervalo $[1, 5)$, hay 14 pacientes que están vivos (en riesgo) antes que $t = 1$ y mueren 3; así, un estimativo de la probabilidad condicional de muerte en este intervalo es $3/14$, por lo que la probabilidad de sobrevivir es $1 - 3/14 = 11/14$. También se puede escribir como:

$$\widehat{S}(1) = \widehat{P}(\mathbf{T} \geq 0)\widehat{P}(\mathbf{T} \geq 1|\mathbf{T} \geq 0) = (1)(11/14) = 0.786.$$

De esta manera, se calculan las estimaciones para los demás intervalos. Estos valores se consignan en la tabla 2.2.

Tabla 2.1. Sobrevida en estudio de hepatitis

Control	1^+	2^+	3	3	3^+	5^+	5^+	16^+
	16^+	16^+	16^+	16^+	16^+	16^+	16^+	
Esteroide	1	1	1	1^+	4^+	5	7	8
	10	10^+	12^+	16^+	16^+	16^+		

Tabla 2.2. Sobrevida en estudio de hepatitis

$t_{(j)}$	Intervalos	$e_{(j)}$	$n_{(j)}$	$\widetilde{S}_{KM}(t)$
0	[0, 1)	0	14	1.000
1	[1, 5)	3	14	0.786
5	[5, 7)	1	9	0.698
7	[7, 8)	1	8	0.611
8	[8, 10)	1	7	0.524
10	[10, 16)	1	6	0.437

2.4. Bandas de confianza para la función de sobrevida

Una ayuda en la interpretación del estimador de un parámetro es la precisión del estimador, la cual se consigue con el error estándar del estimador. El error estándar ($e.e(\cdot)$) se define como la raíz cuadrada de la varianza del estimador. Así, si $\hat{\theta}$ es un estimador del parámetro θ, su error estándar es $e.e(\hat{\theta}) = \sigma_{\hat{\theta}} = \sqrt{var(\hat{\theta})}$. El estimador del parámetro y el respectivo error estándar sirven en la construcción de un intervalo de confianza para el parámetro. Se desarrolla en esta sección la construcción de un intervalo de confianza para la función de sobrevida.

El estimador KM de la función de sobrevida es

$$\tilde{S}_{KM}(t) = \prod_{j:t_{(j)}<t} \left(\frac{n_{(j)} - e_{(j)}}{n_{(j)}} \right) = \prod_{j:t_{(j)}<t} \hat{p}_j, \tag{2.4.1}$$

donde $\hat{p}_j = (n_{(j)} - e_{(j)})/n_{(j)}$ es la probabilidad de que una unidad sobreviva (no le ocurra el evento) en un intervalo que empieza en $t_{(j)}$. Tomando

logaritmos,

$$\ln \tilde{S}_{KM}(t) = \sum_{j=1}^{k} \ln \hat{p}_j,$$

así, la varianza de $\ln \tilde{S}_{KM}(t)$ viene dada por

$$var(\ln \tilde{S}_{KM}(t)) = \sum_{j=1}^{k} var(\ln \hat{p}_j)_{KM}. \tag{2.4.2}$$

El número de unidades que sobreviven (no les ocurre el evento) en un intervalo que empieza en $t_{(j)}$ se puede considerar como una variable aleatoria con distribución binomial con parámetros $n_{(j)}$ y p_j. El número de unidades que sobreviven en este intervalo es $n_{(j)} - e_{(j)}$ y como la varianza de una $v.a.$ binomial es $np(1-p)$, entonces

$$var(n_{(j)} - e_{(j)}) = n_{(j)} p_j (1 - p_j).$$

Ahora, como $\hat{p}_j = (n_{(j)} - e_{(j)})/(n_{(j)})$, entonces

$$var(\hat{p}_j) = var\left(\frac{n_{(j)} - e_{(j)}}{n_{(j)}}\right) = \frac{1}{n_{(j)}^2} var[n_{(j)} - e_{(j)}] = \frac{p_j(1-p_j)}{n_{(j)}}.$$

En consecuencia, una estimación de esta varianza es

$$\widehat{var}(\hat{p}_j) = \frac{\hat{p}_j(1-\hat{p}_j)}{n_{(j)}}.$$

Una de las versiones del *Método Delta* establece que la varianza de la función $g(X)$ de una variable aleatoria X se aproxima mediante la siguiente expresión:

$$var(g(X)) \approx \left(\frac{dg(X)}{dX}\right)^2 var(X). \tag{2.4.3}$$

Para este caso, se tiene que $g(\hat{p}_j) = \ln \hat{p}_j$. Así una aproximación a su varianza es

$$var(\ln \hat{p}_j) \approx \frac{var(\hat{p}_j)}{\hat{p}_j^2} \approx \frac{(1-\hat{p}_j)}{n_{(j)}\hat{p}_j} = \frac{e_{(j)}}{n_{(j)}(n_{(j)} - e_{(j)})}. \tag{2.4.4}$$

Se retoma la expresión (2.4.3) y se aplican sobre esta los resultados anteriores, así:

$$var(\ln \tilde{S}_{KM}(t)) = \frac{1}{\left(\tilde{S}_{KM}(t)\right)^2} var(\tilde{S}_{KM}(t)).$$

De aquí,

$$var(\tilde{S}_{KM}(t)) \approx \left(\tilde{S}_{KM}(t)\right)^2 \sum_{j=1}^{k} \frac{e_{(j)}}{n_{(j)}(n_{(j)} - e_{(j)})}. \tag{2.4.5}$$

Con el resultado anterior, el error estándar del estimador Kaplan-Meier para la función de sobrevivencia es definido por

$$e.e.(\tilde{S}_{KM}(t)) = \left(\tilde{S}_{KM}(t)\right) \left\{ \sum_{j=1}^{k} \frac{e_{(j)}}{n_{(j)}(n_{(j)} - e_{(j)})} \right\}^{\frac{1}{2}}, \tag{2.4.6}$$

para tiempos t tal que $t_{(k)} \leq t < t_{(k+1)}$. Este resultado es conocido como *la fórmula de Greenwood*.

Se pueden obtener intervalos de confianza puntuales para la función de sobrevivencia. Estos intervalos son válidos para un solo valor del tiempo, sobre el cual se hace inferencia.

En algunos análisis resulta interesante obtener bandas de confianza que garanticen, con un nivel dado de confiabilidad, que la función de sobrevivencia esté dentro de la banda para todo t tal que $t_{(k)} \leq t < t_{(k+1)}$; por esto se denomina *banda de confianza*, pues los límites de confianza son constantes dentro del intervalo $[t_{(k)}, t_{(k+1)})$.

Este intervalo de confianza, para el verdadero valor de la función de supervivencia al tiempo t, se obtiene con el supuesto de que el valor estimado de la función de sobrevivencia al tiempo t, $\tilde{S}_{KM}(t)$, se distribuye asintóticamente, conforme a una distribución normal con media $S(t)_{KM}$ y varianza estimada dada por $var(\tilde{S}_{KM}(t))$. Por tanto, un intervalo del $100 \times (1-\alpha)\,\%$ de confianza para $S(t)$, para un valor específico t, es dado por

$$\tilde{S}_{KM}(t) \mp Z_{(1-\alpha/2)} \sqrt{\widehat{var}\tilde{S}_{KM}(t)\}}. \tag{2.4.7}$$

De manera explícita, el intervalo de confianza para $S(t)_{KM}$ es

$$\left(\tilde{S}_{KM}(t) - Z_{(1-\alpha/2)} e.e.\{\tilde{S}(t)\}; \quad \tilde{S}_{KM}(t) + Z_{(1-\alpha/2)} e.e.\{\tilde{S}_{KM}(t)\}\right), \tag{2.4.8}$$

donde $e.e.\{\tilde{S}_{KM}(t)\}$ está dado por la fórmula de Greenwood y $Z_{(1-\alpha/2)}$ es el cuantil que acumula una probabilidad de $(1-\alpha/2)$ en una distribución normal estándar.

No obstante, para valores extremos de t, este intervalo de confianza puede presentar limites inferiores negativos o límites superiores mayores que

1. En tales casos, el problema se resuelve utilizando una transformación para $S_{KM}(t)$ como, por ejemplo, $\tilde{U}(t) = \ln[-\ln(\tilde{S}_{KM}(t))]$ que tiene varianza asintótica estimada por

$$var(\tilde{U}(t)) \approx \frac{1}{\left(\ln \tilde{S}_{KM}(t)\right)^2} \sum_{j=1}^{k} \frac{e_{(j)}}{n_{(j)}(n_{(j)} - e_{(j)})} = \frac{\sum_{j=1}^{k} \frac{e_{(j)}}{n_{(j)}(n_{(j)} - e_{(j)})}}{\left[\sum_{j=1}^{k} \ln\left(\frac{e_{(j)}}{n_{(j)}(n_{(j)} - e_{(j)})}\right)\right]^2}. \tag{2.4.9}$$

De esta manera, un intervalo del $(1 - \alpha)\,\%$ de confianza para $S_{KM}(t)$ es

$$\tilde{S}_{KM}(t)^{\exp\left\{\pm Z_{(1-\alpha/2)}\sqrt{var\tilde{U}(t)}\right\}}, \tag{2.4.10}$$

el cual toma valores en el intervalo $[0, 1]$.

2.5. Otros estimadores no paramétricos

El estimador de Kaplan-Meier es uno de los más utilizados e incorporados en los paquetes estadísticos con los cuales se hacen los cálculos para obtener valores del estimador de la función de sobrevida $S(t)$. No obstante, hay otros estimadores de $S(t)$; ellos son el estimador de Nelson-Aalen y el estimador de tablas de vida. El estimador de Nelson-Aalen tiene propiedades casi similares al de Kaplan-Meier. El de tablas de vida tiene importancia histórica, pues se utilizó en información proveniente de censos demográficos para estimar características asociadas con el tiempo de vida de humanos. Este estimador fue propuesto por demógrafos y actuarios y es utilizado en muestras de tamaño grande [11].

2.5.1. Estimador de Nelson-Aalen

Este es un estimador más reciente que el de Kaplan-Meier y su estimación está basada en la expresión (1.5.4) para $S(t)$ dada por

$$S(t) = \exp\{-H(t)\},$$

donde $H(t)$ es la función de riesgo acumulado. Un estimador para $S(t)$ fue propuesto inicialmente por Nelson [57] y luego retomado por Aalen [1], quien encontró sus propiedades asintóticas mediante procesos de conteo.

Este estimador se denomina en la literatura como el estimador de Nelson-Aalen (en adelante NA) y es de la siguiente forma:

$$\tilde{S}_{NA}(t) = \exp\{-\tilde{H}(t)\}, \tag{2.5.1}$$

donde

$$\tilde{H}(t) = \sum_{j:t_j<t} \left(\frac{e_{(j)}}{n_{(j)}}\right), \tag{2.5.2}$$

con $e_{(j)}$ y $n_{(j)}$ definidos de manera similar a los empleados en el estimador de Kaplan-Meier. La varianza de este estimador es

$$\widehat{var}(\widehat{H}(t)) = \sum_{j=1}^{k} \left(\frac{e_{(j)}}{n_{(j)}^2}\right).$$

Otro estimador para la varianza de $var(\widehat{H}(t))$ propuesto por Klein [38] es

$$\widehat{var}(\widehat{H}(t)) = \sum_{j=1}^{k} \frac{(n_{(j)} - e_{(j)})e_{(j)}}{n_{(j)}^3},$$

el cual tiene un sesgo menor que el anterior.

La varianza de este estimador puede obtenerse mediante

$$var(\tilde{S}_{NA}(t)) = \left(\tilde{S}(t)\right)^2 \sum_{j=1}^{k} \left(\frac{e_{(j)}}{n_{(j)}^2}\right). \tag{2.5.3}$$

El estimador de Nelson-Aalen $S_{NA}(t)$ y el de Kaplan-Meier $S_{KM}(t)$ son muy próximos a $S(t)$, es decir, consistentes. Bohoris [3] muestra que $S_{NA}(t) \geq S_{KM}(t)$ para todo t. Es decir, el estimador de Nelson-Aalen es mayor o igual que el estimador de Kaplan-Meier.

Para los datos de hepatitis, en el grupo de pacientes que reciben el esteroide, el cálculo del valor de la función de sobrevida a la semana 7, de acuerdo con la expresión (2.5.2), es como sigue:

$$\begin{aligned}
\tilde{H}(t=7) &= \sum_{j:t_j<7} \left(\frac{e_{(j)}}{n_{(j)}}\right) \\
&= \frac{0}{14} + \frac{3}{14} + \frac{1}{9} + \frac{1}{8} \\
&= 0.000 + 0.214 + 0.111 + 0.125 \\
&= 0.450.
\end{aligned}$$

Por lo tanto, el valor de la función de sobrevida en el grupo de pacientes tratado con esteroides a la semana 7, por (2.5.1), es

$$\tilde{S}_{NA}(t=7) = \exp\{-\tilde{H}(t=7)\} = \exp\{-0.450\} = 0.637.$$

El error estándar de la función de sobrevida a la semana 7 se calcula mediante la expresión (2.5.3), este es

$$\begin{aligned} ee(\tilde{S}_{NA}(t=7)) &= \sqrt{var(\tilde{S}_{NA}(t=7))} = \left(\tilde{S}(t=7)\right)\sqrt{\sum_{t:t_j<7}\left(\frac{e_{(j)}}{n^2_{(j)}}\right)} \\ &= (0.637)\sqrt{(0/14^2 + 3/14^2 + 1/9^2 + 1/8^2)} \\ &= (0.637)\sqrt{0.0000 + 0.0153 + 0.0123 + 0.0156} \\ &= 0.1326. \end{aligned}$$

La tabla 2.3 contiene todo los resultados relacionados con las estimaciones para la función de sobrevida mediante el estimador de Nelson-Aalen, su errores estándar y los intervalos de confianza para cada punto en los tiempos observados.

Tabla 2.3. Estimación de la sobrevida mediante Nelson-Aalen

t_j	n_j	e_j	$H(t_j)$	$\tilde{S}_{NA}(t_j^-)$	$ee(\tilde{S}_{NA}(t_j^-))$	L_1	L_2
0	14	0	0.000	1.0000	0		
1	14	3	0.214	0.8071	0.0999	0.611	1.000
5	9	1	0.325	0.7222	0.1201	0.487	0.958
7	8	1	0.450	0.6374	0.1326	0.378	0.897
8	7	1	0.593	0.5525	0.1394	0.279	0.825
10	6	1	0.760	0.4677	0.1414	0.191	0.745

Comparando las tablas 2.2 y 2.3 en las columnas de las respectivas funciones de sobrevida empíricas, $\tilde{S}_{KM}$ y $\tilde{S}_{NA}$, se observa que la función de sobrevida estimada mediante Nelson-Aalen es mayor que la estimada mediante Kaplan-Meier; es decir, para todo t, $\tilde{S}_{NA} > \tilde{S}_{KM}$. Esta propiedad se demuestra en Bohoris [3].

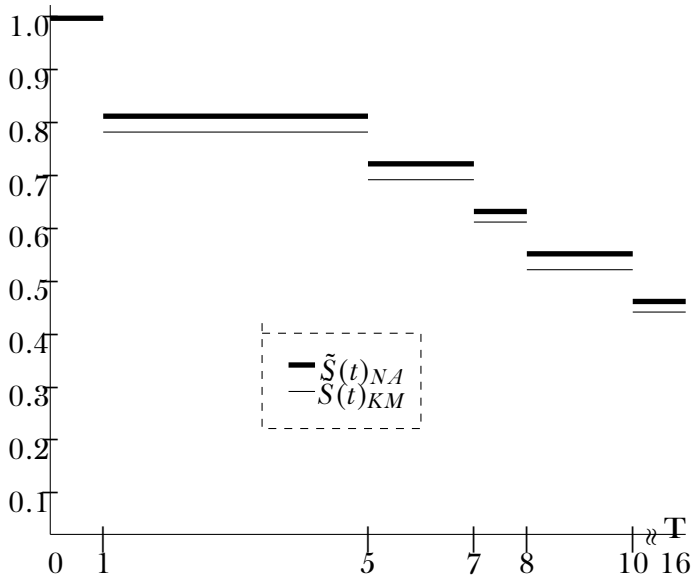

Figura 2.3. Estimadores de KM y NA para la función de sobrevida.

En la gráfica 2.3 se muestran las funciones de sobrevida estimadas mediante Kaplan-Meier (línea delgada) y Nelson-Aalen, respectivamente. Se evidencia que las estimaciones obtenidas mediante Nelson-Aalen son mayores o iguales que las obtenidas por el estimador de Kaplan-Meier.

2.6. Estimador de $h(t)$ mediante el suavizador de núcleo

En la mayoría de los estudios, el interés no es la función de riesgo acumulado, $H(t)$, sino su derivada, $h(t)$, que es la tasa de riesgo. Por la definición de $H(t)$, la pendiente del estimador de Nelson-Aalen suministra un estimador empírico de la tasa de riesgo $h(t)$. En esta sección, la estimación de $h(t)$ se hace mediante un suavizador tipo núcleo (en inglés, *kernel*).

Los estimadores de $h(t)$ tipo suavizadores de núcleo se basan en el estimador de Nelson-Aalen de la función de riesgo acumulado, $\tilde{H}(t)$. Este estimador puede construirse también para datos censurados a la derecha o para datos truncados a la izquierda.

La función $\tilde{H}(t)$ es una función escalonada con saltos en los tiempos donde se presenta el evento, $0 = t_0 < t_1 < \cdots < t_r$. Las cantidades $\Delta\tilde{H}(t_i) = \tilde{H}(t_i) - \tilde{H}(t_{i-1})$ y $\Delta\widehat{Var}(\tilde{H}(t_i)) = \widehat{Var}(\tilde{H}(t_i)) - \widehat{Var}(\tilde{H}(t_{i-1}))$

son la magnitud de los saltos de la función de riesgo acumulada estimada, $\tilde{H}(t_i)$, y su varianza, $\widehat{Var}(\tilde{H}(t_i))$, en el tiempo t_i, respectivamente. Se puede advertir que $\Delta\tilde{H}(t_i)$ es un estimador empírico de la función $h(t)$ en los tiempos de ocurrencia del evento.

El estimador de $h(t)$ mediante un suavizador de núcleo es un promedio ponderado de estos estimadores empíricos sobre los tiempos del evento cercanos a T. La "cercanía" es determinada por un radio de la vecindad centrada en t con un ancho de banda, b, tal que los tiempos al evento en el rango $t - b$ a $t + b$ sean incluidos en el promedio ponderado que estima $h(t)$. El ancho de banda, b, es escogido de manera que minimice el error cuadrático medio en algún grado de suavizamiento establecido. Las ponderaciones son controladas por la escogencia de una función núcleo, $K(\cdot)$, definida en el intervalo $[-1, +1]$, la cual determina el valor del peso de un punto a una distancia de t.

Las funciones núcleo más frecuentes son:

- Uniforme: $K(x) = \frac{1}{2}$, para $-1 \leq x \leq 1$.
- Epanechnikov: $K(x) = 0.75(1 - x^2)$, para $-1 \leq x \leq 1$.
- Biponderado: $K(x) = \frac{15}{16}(1 - x^2)^2$, para $-1 \leq x \leq 1$.
- Triangular: $1 - |x|$, para $-1 \leq x \leq 1$.
- Gausiano: $\frac{1}{\sqrt{2\pi}} e^{-(1/2)x^2}$, para $-\infty < x < \infty$.
- Arco coseno: $\frac{\pi}{4} \cos(\pi x/2)$, para $-1 \leq x \leq 1$.

El núcleo uniforme da el mismo peso a todas los eventos ocurridos dentro del intervalo $[t - b, t + b]$, mientras que en los otros dos el peso aumenta en tanto el evento ocurra cerca a t.

El estimador de la tasa de riesgo mediante el núcleo es definida para todos los puntos de tiempo $t > 0$. Para todos los puntos t tales que $b \leq t \leq t_r - b$, el estimador tipo suavizador de núcleo de $h(t)$, basado en la función núcleo $K(\cdot)$, es dado por:

$$\hat{h}_{ker}(t) = \frac{1}{b} \sum_{i=1}^{r} K\left(\frac{t - ti}{b}\right) \Delta(\widehat{H}(t_i)). \tag{2.6.1}$$

Se mantiene la denominación "*ker*" para indicar el núcleo.

La varianza de $\hat{h}_{ker}(t)$ es estimada por

$$\widehat{Var}(\hat{h}_{ker}(t)) = \frac{1}{b^2} \sum_{i=1}^{r} K\left(\frac{t - ti}{b}\right)^2 \Delta\widehat{Var}\widehat{H}(t_i). \tag{2.6.2}$$

Los intervalos o bandas de confianza para la tasa de riesgo estimada, por suavizamiento tipo núcleo, se pueden construir de manera similar a las construidas antes para la función de sobrevida estimada mediante Kaplan-Meier o Nelson-Aalen.

Mediante R, con el paquete *muhaz*, se calcula y gráfica la función tasa de riesgo a partir de datos censurados a la derecha utilizando los métodos basados en el suavizamiento por núcleos. Las opciones incluyen tres tipos de funciones para el ancho de banda, tres tipos de corrección de límite y cuatro formas para la función núcleo.

Ejemplo 2.6.1. En un estudio sobre cáncer de ovario, citado en Therneau y Grambsch [66], en Collet [10] y en la documentación del paquete *muhaz* de R, se hizo seguimiento a 26 mujeres con un mínimo de residuos de la enfermedad y que participaban de un tratamiento de quimioterapia. En el estudio se aplicaron dos tipos de quimioterapia sobre dos grupos de mujeres. La variable respuesta fue el tiempo de sobrevida en los días siguientes al inicio del tratamiento de quimioterapia asignado aleatoriamente. A continuación, se reseñan las variables cuyos datos están en la tabla 2.4.

- Tiempo: tiempo de sobrevida en días.
- Censura: indicador del evento (0 = censura, 1 = no censura).
- Edad: edad del paciente en años.
- Residuos Enf.: residuos de la enfermedad (1 = no, 2 = si).
- Trat.: tratamiento (1 = simple, 2 = combinado).
- Estado: estado (1 = bueno, 2 = malo).

Con el siguiente código de R se obtiene la gráfica suavizada mediante núcleo de .[Ep]anechnikov"(por defecto). El ancho de banda es el mismo para todos los puntos de grilla.

```
#Leer los datos de la Tabla 2.4
can_ovario<-read.table("Cancer_ovario.txt",header=T)
attach(can_ovario)
# Cargar los paquetes survival y muhaz
require(survival)
require(muhaz)
# Estimaci\'on de la funci\'on de riesgo mediante n\'ucleos
riesgo.can<- muhaz(Tiempo, Censura)
plot(riesgo.can)
```

Tabla 2.4. Tiempos de sobrevida de mujeres con cáncer de ovario

	Tiempo	Censura	Edad	Residuos Enf.ds	Trat.	Estado
1	59	1	72.3315	2	1	1
2	115	1	74.4932	2	1	1
3	156	1	66.4658	2	1	2
4	421	0	53.3644	2	2	1
5	431	1	50.3397	2	1	1
6	448	0	56.4301	1	1	2
7	464	1	56.9370	2	2	2
8	475	1	59.8548	2	2	2
9	477	0	64.1753	2	1	1
10	563	1	55.1781	1	2	2
11	638	1	56.7562	1	1	2
12	744	0	50.1096	1	2	1
13	769	0	59.6301	2	2	2
14	770	0	57.0521	2	2	1
15	803	0	39.2712	1	1	1
16	855	0	43.1233	1	1	2
17	1040	0	38.8932	2	1	2
18	1106	0	44.6000	1	1	1
19	1129	0	53.9068	1	2	1
20	1206	0	44.2055	2	2	1
21	1227	0	59.5890	1	2	2
22	268	1	74.5041	2	1	2
23	329	1	43.1370	2	1	1
24	353	1	63.2192	1	2	2
25	365	1	64.4247	2	2	1
26	377	0	58.3096	1	2	1

La gráfica que resulta del procesamiento con el paquete *muhaz* se muestra en la figura 2.4. El riesgo a la muerte en mujeres con cáncer de ovario, después de iniciar un tratamiento de quimioterapia, como se muestra en la figura 2.4, crece con una tasa constante hasta aproximadamente el primer año (365 días) de iniciado el tratamiento. Después de este tiempo el riesgo disminuye y después de año y medio el riesgo a la muerte cae rápidamente.

2.6.1. Estimador de tabla de vida actuarial

La construcción de una tabla actuarial de vida consiste en dividir el periodo de tiempo en un cierto número de intervalos. Suponga que el eje del tiempo se divide en $s+1$ intervalos, los cuales están determinados por los siguientes

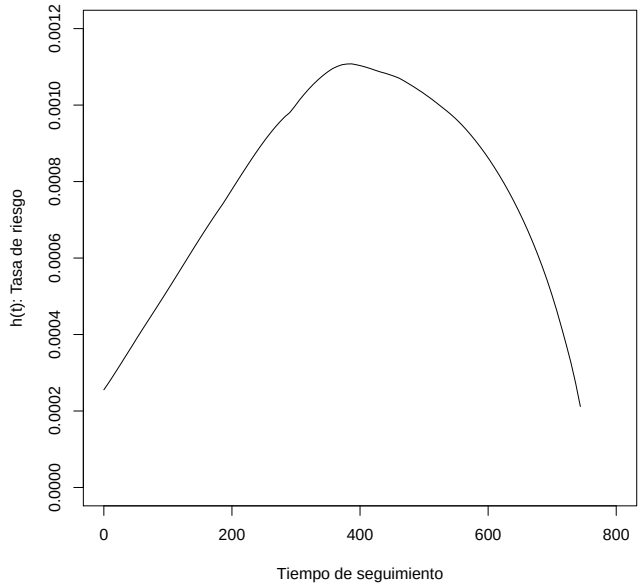

Figura 2.4. Función de riesgo de tiempo de sobrevida a cáncer de ovario, suavizada mediante núcleo.

puntos de corte, $t_1, t_2, \ldots, t_s$. Así, el j-ésimo intervalo, $I_j = [t_{j-1}; t_j)$; $j = 1, \ldots, s$, donde $t_0 = 0$ y $t_s = +\infty$.

Un estimador para la tabla de vida se obtiene desde el estimador de Kaplan-Meier, con una modificación de la probabilidad $q_{(j)}$ contenida en la expresión (2.3.2). Así,

- $n_{(j)} = \{\text{Núm. de eventos en el intervalo } [t_{j-1}; t_j)\}$,
- $e_{(j)} = \{\text{Núm. de unidades al tiempo, } t_{j-1}\} - \frac{1}{2}\{\text{Núm. de censuras en } I_j\}$.

Una adecuación para $q_{(j)}$ en tablas de vida es

$$\tilde{q}_{(j)} = \frac{\{\text{Núm. de eventos en el intervalo } [t_{j-1}; t_j)\}}{\{\text{Núm. de unidades al tiempo } t_{j-1}\} - \frac{1}{2}\{\text{Núm. de censuras en } I_j\}}.$$

Reemplazando esta expresión para $\tilde{q}_{(j)}$ en (2.3.2) se obtiene que el estimador para $S(t)$ en una tabla de vida es

$$\tilde{S}(t \in I_j) = \begin{cases} 1, & \text{para } j = 1 \\ \tilde{S}(t \in I_{j-1}), & \text{para } j = 2, \ldots, s. \end{cases}$$

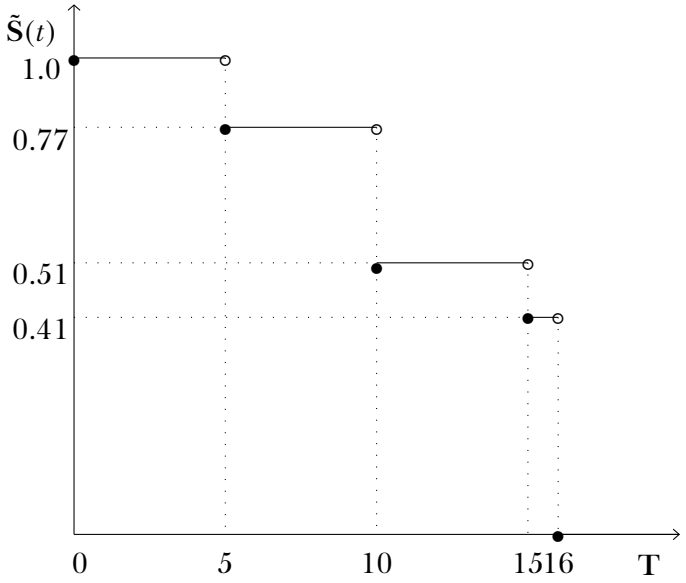

Figura 2.5. Función de sobrevida en estudio de hepatitis.

De otra manera,

$$\tilde{S}(t) = \prod_{i=2}^{j}(1 - \tilde{q}_{i-1}),\ t \in I_j, \tag{2.6.3}$$

para $j = 1, \ldots, s$ y $\tilde{q}_0 = 0$. La representación gráfica de esta función de sobrevida es una función escalonada no creciente y constante en cada intervalo de tiempo.

Para los datos de la tabla 2.1 sobre el grupo de pacientes tratados con un esteroide, se dividen los tiempos en 4 intervalos: $[0, 5)$, $[5, 10)$, $[10, 15)$ y $[15, 16)$. Una estimación de q_2 correspondiente al intervalo $I_2 = [5, 10)$ es $\tilde{q}_2 = 3/9 = 0.33$. Esto significa que la probabilidad de muerte antes de la semana 10 para quienes sobrevivieron a la semana 5 es 33 %. La tabla 2.3 tiene los demás resultados para cálculos similares en los otros intervalos.

Una estimación de la función de sobrevida en el tiempo $t = 10$ semanas es:

$$\tilde{S}(10) = (1 - 0.231)(1 - 0.333) = 0.513.$$

Se interpreta que una persona con hepatitis tratada con esteroide tiene una probabilidad de 51.3 % de sobrevivir a la semana 10 de tratamiento. La tabla 2.3 contiene los valores estimados de la función de sobrevida para los demás intervalos.

Tabla 2.5. Sobrevida en estudio de hepatitis

Intervalo	Pac/riesgo	Muertes	Censuras			
I_j	$n_{(j)}$	$e_{(j)}$	$c_{(j)}$	$\tilde{q}_j$	$(1-\tilde{q}_j)$	$\tilde{S}(t)$
$\widetilde{S}_{KM}(t)$						
$[0, 5)$	14	3	2	0.231	0.769	1.0
$[5, 10)$	9	3	0	0.333	0.667	0.769
$[10, 15)$	6	1	2	0.200	0.800	0.513
$[15, 16)$	3	0	3	0.000	1.000	0.410

En la figura 2.5 se muestra la función de sobrevida para los datos de la tabla 2.3. Se advierte que es una función escalonada no creciente, con continuidad a la derecha.

2.7. Comparación de funciones de sobrevida

El objetivo de algunos estudios puede dirigirse a la comparación de la sobrevida entre varios grupos de unidades en un determinado periodo de tiempo. Los grupos pueden ser inducidos por las unidades que reciben un tratamiento o por los diferentes valores que toma una variable (categórica o categorizada). Así, por ejemplo, en medicina se quiere comparar la sobrevida de los grupos de pacientes que reciben tres medicamentos diferentes, los cuales son suministrados para el tratamiento de una determinada dolencia; en el caso del tiempo transcurrido para que un niño aprenda a leer (o escribir), los grupos pueden corresponder al sexo (niños o niñas), a las condiciones sociodemográficas de los niños, a las prácticas pedagógicas o a los grupos de edad. Así, los grupos pueden ser naturales o inducidos por un tratamiento o categorización de una variable como la edad.

La primera herramienta para hacer una comparación entre las curvas de sobrevida de los respectivos grupos es la elaboración de las gráficas de las funciones de sobrevida, las cuales deben ser estimadas sobre el mismo eje de tiempo. No obstante, la comparación de las funciones de sobrevida no siempre es evidente desde un punto de vista gráfico, pues todo gráfico es finito (en un rango de valores de **T**), por lo que habrán características que un gráfico no revela al lector. Es decir, si dos funciones de sobrevida S_1 y S_2 son tales que para algún tiempo t $S_1(t) > S_2(t)$ en un rango determinado,

esta desigualdad puede cambiar en otros rangos de la variable **T**, debido a posibles coincidencias, cruces o traslapes de las gráficas asociadas con las funciones de sobrevida.

La figura 2.6 muestra dos posibles escenarios de funciones de sobrevida. En la figura del panel a, la función de sobrevida $S_1(t)$ es mayor que la función de sobrevida $S_2(t)$ en todo el rango de tiempo observable (0 a 10 unidades de tiempo). En el panel b, la diferencia entre las dos funciones de sobrevida no es "uniforme", pues en el rango de tiempo 0 a 3 se tiene que $S_1(t) > S_2(t)$, mientras que en el rango 3 a 10 la desigualdad se invierte. De esta manera, se muestra que el análisis gráfico no es concluyente o confirmatorio.

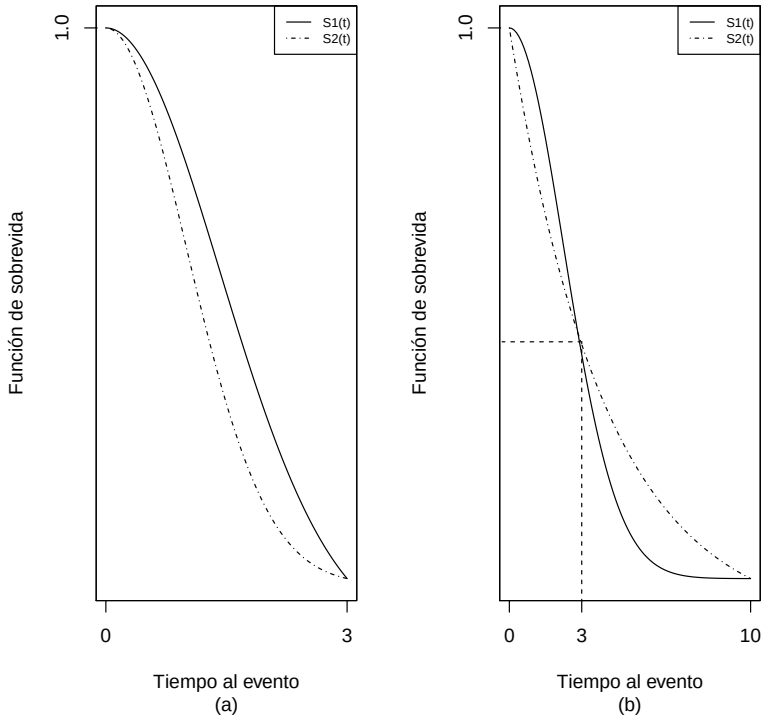

Figura 2.6. Comparación de funciones de sobrevida.

En esta sección se presentan algunos procedimientos alternativos para verificar la existencia de una posible diferencia entre las funciones de sobrevida junto con las estadísticas no paramétricas tales como *log-rank*, *Wilcoxon* y *Tarone-Ware*.

2.7.1. Estadística log-rank

Esta estadística es apropiada cuando la razón de las funciones de sobrevida de los grupos en riesgo al evento es aproximadamente constante. Es decir, los grupos tienen riesgos al evento proporcionales entre sí.

Se considera la prueba de igualdad de dos funciones de sobrevida $S_1(t)$ y $S_2(t)$, es decir, $S_1(t) = S_2(t)$, asociadas con los grupos G_1 y G_2, respectivamente. Sean $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ los tiempos al evento de la muestra conformada por la combinación de las dos muestras de tamaños n_1 y n_2. Suponga que en el tiempo $t_{(j)}$ ocurren $e_{(j)}$ eventos y $n_j = n_{1j} + n_{2j}$ unidades expuestas al evento (en riesgo) en un tiempo inmediatamente inferior a $t_{(j)}$ de la muestra combinada, y e_{jg} y n_{jg} en la muestra $g = 1, 2$ y $j = 1, \ldots, k$. En cada tiempo al evento $t_{(j)}$, los datos pueden disponerse en una tabla de contingencia 2×2 con e_{jg} eventos y $n_{jg} - e_{jg}$ unidades en riesgo al evento en la columna g–ésima.

Tabla 2.6. Núm. de eventos por grupo al momento $t_{(j)}$

	Grupos		
Evento (**E**)	1	2	Total
Ocurre	e_{j1}	e_{j2}	e_j
No ocurre	$n_{j1} - e_{j1}$	$n_{j2} - e_{j2}$	$n_j - e_j$
Total	n_{j1}	n_{j2}	n_j

Condicionando la ocurrencia del evento y la censura al tiempo $t_{(j)}$ (fijando los márgenes por columna) y el número de eventos (fallas) en el tiempo $t_{(j)}$ (fijando los márgenes por fila), la distribución de e_{j2} es una hipergeométrica:

$$P(e_{j2}) = \frac{\binom{n_{j1}}{e_{j1}}\binom{n_{j2}}{e_{j2}}}{\binom{n_j}{e_j}}. \tag{2.7.1}$$

La media de e_{2j} es $\mu_{e_{j2}} = n_{j2}e_j/n_j$, lo cual equivale a decir que, si no hay diferencia entre los dos grupos respecto a la sobrevida al tiempo $t_{(j)}$, el número total de eventos (e_j) puede dividirse entre las dos muestras de acuerdo con la razón entre el número de unidades expuestas (en riesgo) en cada muestra y el número total de unidades expuestas. La varianza de e_{j2}, obtenida a

partir de la distribución hipergeométrica, es

$$\sigma^2_{e_{j2}} = \frac{n_{j1} n_{j2} e_j (n_j - e_j)}{n_j^2 (n_j - 1)}.$$

Así, la estadística $e_{j2} - \mu_{e_{j2}}$ tiene media cero y varianza $\sigma^2_{e_{j2}}$. Si las k tablas de contingencia son independientes, una estadística aproximada para verificar la igualdad de las dos funciones de sobrevida puede hacerse mediante la estadística

$$LR = \frac{\sum_{j=1}^{k} \left[e_{j2} - \mu_{e_{j2}} \right]^2}{\sum_{j=1}^{k} \sigma^2_{e_{j2}}}, \tag{2.7.2}$$

la cual, para muestras de tamaño grande, tiene distribución ji-cuadrado con 1 grado de libertad.

Esta estadística combina la información contenida en una tabla 2×2 como la anterior y se basa en la estadística de Mantel-Haenszel como se muestra en Díaz, Morales y León [22]. De ahí que tome los nombres de Mantel-Cox y Peto-Mantel-Haenszel, pero el más conocido es de *log-rank*. La explicación de este nombre es que la estadística puede obtenerse a partir de los rangos de los tiempos al evento en los dos grupos, y la estadística resultante de rangos se basa en el logaritmo del estimador de Nelson-Aalen de la función de sobrevida.

La estadística resume la extensión de desvíos entre los datos de tiempo para el evento y su valor esperado, bajo la hipótesis nula de no diferencia entre estos. Los valores grandes de la estadística LR (equivalentes a valores–p pequeños) advierten sobre la existencia de posibles diferencias entre las funciones de sobrevida de los dos grupos.

La estadística log-rank puede extenderse para verificar la igualdad entre las funciones de sobrevida, $S_1(t), \ldots, S_q(t)$, asociadas a cada uno de q grupos, respectivamente. A la notación anterior se adiciona el subíndice $g = 1, \ldots, q$, para señalar el grupo al que pertenece la función de sobrevida. Así, los datos pueden ser dispuestos en una tabla de contingencia de tamaño $2 \times q$ con d_{jg} unidades a las cuales les ha ocurrido el evento y $n_{jg} - d_{jg}$ unidades expuestas al evento en la columna $g = 1, \ldots, q$. La tabla de datos tiene la forma mostrada en la tabla 2.7. De manera análoga al caso de dos grupos, condicionando la ocurrencia del evento y la censura al tiempo t_j (fijando los márgenes por columna) y el número de eventos (fallas) en el tiempo t_j (fijando los márgenes por fila), la distribución conjunta de $d_{j1}, d_{j2}, \ldots, d_{jq}$ es

Tabla 2.7. Extensión a g-grupos del evento al momento $t_{(j)}$

	Grupos					
Evento (E)	1	$\cdots$	g	$\cdots$	q	Total
Ocurre	e_{j1}	$\cdots$	e_{jg}	$\cdots$	e_{jq}	e_j
No ocurre	$n_{j1} - e_{j1}$	$\cdots$	$n_{jg} - e_{jg}$	$\cdots$	$n_{jq} - e_{jq}$	$n_j - e_j$
Total	n_{j1}	$\cdots$	n_{jg}	$\cdots$	n_{jq}	n_j

una distribución hipergeométrica multivariada, esta es

$$P(e_{j1}, ed_{j2}, \ldots, ed_{jq}) = \frac{\prod_{g=1}^{q} \binom{n_{jg}}{e_{jg}}}{\binom{n_j}{e_j}}. \tag{2.7.3}$$

La media de e_{jg} es $\mu_{e_{jg}} = n_{jg}\left(e_j/n_j\right)$, la varianza de e_{jg} y la covarianza entre e_{jg} y e_{jh} son, respectivamente,

$$\sigma^2_{e_{jg}} = \frac{n_{jg}(n_j - n_{jg})e_j((n_j - e_j)}{n_j^2(n_j - 1)}$$

y

$$\sigma_{j(gh)} = -\frac{n_{jg}n_{jh}e_j((n_j - e_j)}{n_j^2(n_j - 1)}.$$

La estadística $E'_j = (e_{j1} - \mu_{e_{j1}}, \ldots, e_{jq} - \mu_{e_{jq}})$ tiene media cero y matriz de varianzas-covarianzas $\Sigma_j = \left(\sigma_{j(gh)}\right)$. Se conforma la estadística $\mathscr{E}$ sumando sobre todos los tiempos de falla

$$\mathscr{E} = \sum_{j=1}^{k} E_j,$$

donde $\mathscr{E}$ es un vector de tamaño $q \times 1$ el cual contiene la discrepancia entre los totales de eventos observados y esperados al momento t_j.

Bajo el supuesto de que las k tablas de contingencia son independientes, la matriz de varianzas-covarianzas del vector $\mathscr{E}$ es $\Sigma = \Sigma_1 + \cdots + \Sigma_k$. Una estadística para la verificación de la igualdad de las q funciones de sobrevida se basa en la estadística

$$\mathscr{LR} = \mathscr{E}'\Sigma^{-1}\mathscr{E}, \tag{2.7.4}$$

esta forma cuadrática tiene, para muestras de tamaño grande, una distribución ji-cuadrado con $q - 1$ grados de libertad.

Si se rechaza $H_0 : S_1 = \cdots = S_q$, el siguiente paso es explorar y confirmar cuál o cuáles son las funciones de sobrevida que difieren significativamente de las demás, e incluso ordenarlas a la manera de comparaciones múltiples del análisis de varianza. Una primera herramienta, si el número de grupos, q, es pequeño, es hacer comparaciones dos a dos mediante la estadística log-rank, controlando el error Tipo I mediante la desigualdad de Bonferroni.

Ejemplo 2.7.1. Se quiere comparar tres tipos de condiciones ambientales bajo las cuales germina una misma variedad de semillas de fríjol. Un primer ambiente ($A1$) es el natural, el segundo ambiente ($A2$), bajo invernadero, tiene humedad, temperatura y brillo solar en un nivel bajo, mientras que el tercer ambiente ($A3$) tiene estas condiciones ambientales en un nivel alto. El propósito es evidenciar y detectar si hay diferencia significativa entre los tiempos (en días) hasta la germinación de estas semillas. Las mediciones se registraron durante 12 días a partir de la siembra. La tabla 2.8 contiene los datos en forma creciente por ambiente, el símbolo + significa que es un dato censurado a la derecha.

El objetivo es comparar los tres ambientes de cultivo para evaluar la favorabilidad o eficacia a la germinación temprana.

Tabla 2.8. Tiempo a la germinación en semillas de fríjol

Ambiente 1	5	6	7	7	7	8	8	8
	9	9	10	11	11	12	12	12^+
Ambiente 2	2	3	3	3	5	5	5	6
	6	7	7	8	9	10		
Ambiente 3	5	6	7	7	8	8	8	9
	9	10	10	11	12^+			

Las curvas de sobrevida se estiman mediante el estimador KM y se muestran en la figura 2.7. Se observa que las curvas difieren entre ellas, por lo que se puede hipotetizar que los ambientes de siembra difieren respecto al tiempo que las semillas emplean para su germinación; más aún, en el ambiente 2 (A2) se observa que las semillas tienen una germinación más precoz que en los otros dos ambientes.

El valor de la estadística log-rank, $\mathscr{LR}$ (2.7.4), que bajo la hipótesis de igualdad de las tres curvas de sobrevida tiene una distribución ji-cuadrado con 2 grados de libertad, de acuerdo con los cálculos (del paquete R) es $\mathscr{LR} = 13.7$, con un valor-p igual a 0.00105, lo que puede indicar la existencia de diferencias entre los ambientes de cultivo.

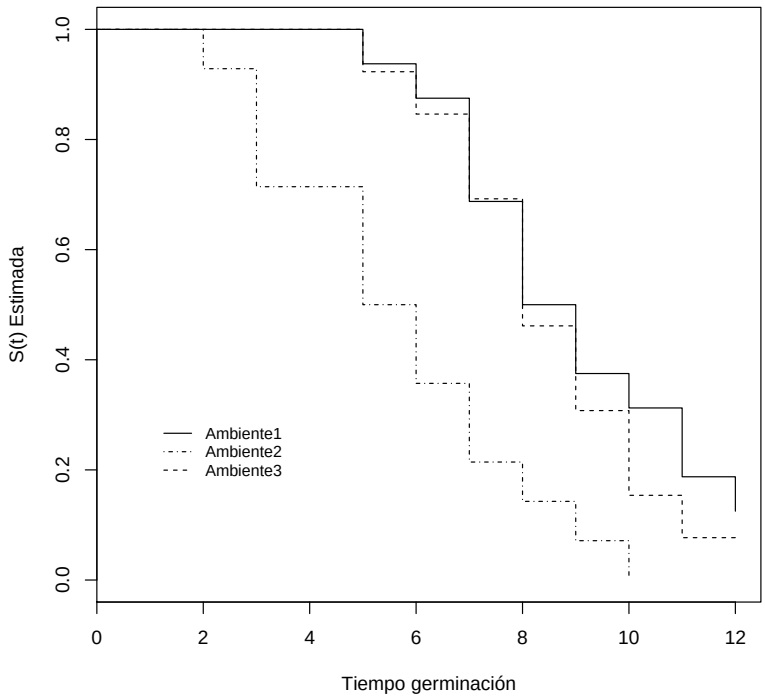

Figura 2.7. Funciones de sobrevida: germinación fríjol.

En el análisis de varianza, una vez que se ha rechazado la hipótesis nula de igualdad de las medias asociadas con q poblaciones, se procede a la búsqueda de las poblaciones (generalmente ligadas a tratamientos) que provocan el rechazo de la hipótesis nula (comparaciones múltiples). Para el caso de comparación de funciones de sobrevida, se trata de establecer cuáles son los grupos o poblaciones que tienen funciones de sobrevida significativamente diferentes y, en lo posible, hacer una jerarquización de estas. Para hacer comparaciones dos a dos entre las tres funciones de sobrevida hay tres posibilidades. Para hacer estos contrastes debe controlarse la inflación del error Tipo I, lo cual se logra mediante la desigualdad de Bonferroni.

Como se pueden hacer tres comparaciones, cada una se hace con un nivel de significancia igual $\alpha/3$; para $\alpha = 0.05$ la significancia de cada comparación es máximo 0.017. De esta forma se garantiza un error Tipo I global con una probabilidad de, como máximo, $0.054 \equiv 5\,\%$. Los resultados de las pruebas log-rank para las comparaciones de funciones de sobrevida se muestran en la tabla 2.9.

Tabla 2.9. Comparación de curvas de sobrevida: datos de fríjol

Ambiente de cultivo	Estadística log-rank	Valor–p
A1 vs A2	12.0	0.00245
A2 vs A3	7.2	0.00747
A1 vs A3	0.1	0.72700

Datos simulados por el autor.

Los datos evidencian que hay diferencias significativas entre los ambientes 1 y 2, 2 y 3. Entre los ambientes 1 y 3 no se muestran diferencias significativas como se observa en la gráfica 2.7. Se puede concluir que el ambiente 2 (A2) es el que acelera el proceso de germinación de estas semillas de fríjol.

Los cálculos y gráficos se hacen mediante el paquete computacional R, el código y la sintaxis para este ejemplo son los que siguen:

```
> # Entrada de los datos
> t<-c(5,6,7,7,7,8,8,8,9,9,10,11,11,12,12,12, 2,3,3,3,5,5,5,6,6,7,7,8,
+ 9,10,5,6,7,7,8,8,8,9,9,10,10,11,12)
> cens<-c(rep(1,15),0,rep(1,14),rep(1,12),0)
> grupo<-c(rep(1,16),rep(2,14),rep(3,13))
> require(survival) # Carga el paquete survival
> ekm<-survfit(Surv(t,cens)~grupo)
> summary(ekm)
> plot(ekm,lty=c(1,4,2),xlab="Tiempo germinaci\'on",ylab="S(t) Estimada")
> legend(1,0.3,lty=c(1,4,2),c("Ambiente1","Ambiente2","Ambiente3"),
+ lwd=1,bty="n",cex=0.8)
> #Compara los tres grupos
> survdiff(Surv(t,cens)~grupo,rho=0)
> #Compara de a dos grupos
> survdiff(Surv(t[1:31],cens[1:31])~grupo[1:31],rho=0)  #Grupo 1vs2
> survdiff(Surv(t[17:44],cens[17:44])~grupo[17:44],rho=0) #Grupo 2vs3
>survdiff(Surv(c(t[1:16],t[32:44]),c(cens[1:16],cens[32:44]))~
+c(grupo[1:16],grupo[32:44]),rho=0) #Grupo 1vs3
```

2.7.2. Otras estadísticas (Wilcoxon y Tarone-Ware)

Una estadística que generaliza la estadística log-rank y que es empleada también para verificar la hipótesis de que no hay diferencia entre las funciones de sobrevida asociadas al tiempo para un evento en dos grupos de unidades es

$$\mathcal{G} = \frac{\left[\sum_{j=1}^{k} w_j(e_{j2} - \mu_{e_{j2}})\right]^2}{\sum_{j=1}^{k} w_j^2 \sigma_{e_{j2}}^2}, \tag{2.7.5}$$

donde w_j son las ponderaciones que determinan diferentes estadísticas. Bajo el supuesto de que las funciones de sobrevida son iguales, la estadística $\mathcal{G}$, para muestras grandes, tiene distribución ji-cuadrado con 1 grado de libertad.

La estadística log-rank se obtiene de la expresión (2.6.4) haciendo las ponderaciones $w_j = 1$. La estadística de *Wilcoxon* se consigue si se hace $w_j = n_j$; la cual se adapta para datos censurados y se basa en la estadística no paramétrica de Wilcoxon empleada para verificar la diferencia entre dos medias poblacionales. Al considerar una ponderación $w_j = \sqrt{n_j}$, es decir, que oscila entre las ponderaciones de log-rank y Wilcoxon, se consigue la estadística de *Tarone-Ware*.

La selección del peso w_j en (2.7.5) pone una dirección al tipo de diferencia que se quiere detectar en las funciones de sobrevida. La estadística de Wilcoxon da un peso igual al número de unidades bajo riesgo al evento. Naturalmente, el peso va disminuyendo en la medida en que transcurre el estudio; el número de unidades decrece en la medida que el evento ocurre o se presentan censuras. En la estadística log-rank, el peso es el mismo a través del tiempo, mientras que la estadística de Tarone-Ware asume un peso intermedio al de log-rank y Wilcoxon.

Una función de la ponderación es una modificación del estimador de Kaplan-Meier, definido de tal forma que sea conocido antes de que ocurra el evento. Un estimador modificado del EKM es

$$\tilde{S}(t) = \prod_{j:t_{(j)}<t} \left(\frac{n_{(j)} + 1 - e_{(j)}}{n_{(j)} + 1}\right). \tag{2.7.6}$$

Las ponderaciones son

$$w_j = \tilde{S}(t_{j-1})\frac{n_{(j)}}{n_{(j)} + 1}.$$

De esta manera, si $\rho = 0$, entonces $w_j = 1$, de donde se obtiene la estadística log-rank. Mientras que si $\rho = 1$, la ponderación es la de KM en el tiempo

anterior al evento, por lo que se obtiene así una estadística similar a la de Wilcoxon.

Este estimador se conoce como el estimador de Peto-Prentice (Peto y Prentice). Una modificación de las ponderaciones contenidas en (2.7.6), propuesta por Harrington, es

$$w_j = \left(\tilde{S}_{KM}(t_{j-1})\right)^{\rho}. \tag{2.7.7}$$

La comparación entre las curvas de sobrevida de los datos relacionados con hepatitis, contenidos en la tabla 2.1 a través de las tres estadísticas anteriores, se muestra en la tabla 2.10. Los cálculos se hacen mediante la opción "survdiff" del paquete *survival* de R con diferentes opciones para el parámetro "rho".

Tabla 2.10. Curvas de sobrevida en datos de hepatitis

Prueba	Estadística	Valor–p
Log-rank	3.67	0.055
Wilcoxon	3.19	0.074
Tarone-Ware	3.43	0.074

De acuerdo con el valor de la ponderación en la estadística (2.7.5), se establece el tipo de diferencia a detectarse en las funciones de sobrevida. La estadística de Wilcoxon utiliza un peso igual al número de individuos en riesgo, por lo que pone más peso o importancia en la parte inicial del tiempo. La estadística de log-rank asigna el mismo peso a todos las unidades en el eje del tiempo. La estadística de Tarone-Ware se encuentra en una situación intermedia.

Capítulo

tres

Modelos probabilísticos en análisis de sobrevida

$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$
$S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) =$
$F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t) = 1 - F(t) = e^{-\theta t}$ $S(t$

3.1. Introducción

La variable *tiempo hasta la ocurrencia del evento*, **T**, es considerada como una variable aleatoria; de manera que esta puede ser de tipo discreto, de tipo continuo o mixta [37, pp. 6-10]. El problema es determinar el modelo probabilístico asociado con la variable aleatoria *tiempo hasta la ocurrencia del evento* para un conjunto de datos en particular.

Seleccionar o proponer la distribución teórica que mejor se aproxime (ajuste) a la distribución natural de los datos (según cada marco conceptual) de los datos es tanto una tarea artística como científica. Es artística porque la postulación de un modelo probabilístico que "genere" los datos requiere del conocimiento empírico y subjetivo que se tenga sobre la variable *tiempo hasta el evento*. Y es científica porque el modelo probabilístico debe ser una auténtica medida de probabilidad. Con esto se anticipa una tarea estadística, que es la verificación del "buen" ajuste del modelo probabilístico a los datos. Mediante gráficos tipo $Q \times Q$ se puede, de manera exploratoria, asegurar el ajuste (de sobrevida) a un modelo probabilístico; en la última sección de este capítulo se describen algunos de esos procedimientos que dan cuenta de la calidad del ajuste de un modelo probabilístico a un conjunto de datos.

En este capítulo se presentan algunos de los modelos probabilísticos (distribuciones de probabilidad) que han sido empleados ampliamente en la descripción probabilística de la variable aleatoria *tiempo hasta un evento*. Para cada uno de los modelos se muestra su identificación, las características distribucionales más importantes y, de acuerdo con las identidades (1.5.2) y (1.5.4), las respectivas funciones de sobrevida y de riesgo.

3.2. Distribución exponencial

Hay muchas situaciones en las que la variable *tiempo para un evento* se ajusta bien a una distribución exponencial, en particular, cuando la variable mide la duración de un dispositivo mecánico o electrónico (análisis de confiabilidad). La distribución exponencial tiene un papel importante en estudios del tiempo hasta el evento (tiempos de vida), de una manera equivalente al papel que tiene la distribución normal en muchas técnicas estadísticas.

Cuando la variable aleatoria *tiempo para el evento* sigue una distribución exponencial con parámetro θ, la función de densidad de probabilidad es definida como

$$f(t) = \begin{cases} \theta e^{-\theta t}, & t \geq 0,\ \theta > 0 \\ 0, & t < 0. \end{cases} \tag{3.2.1}$$

Para indicar que la variable aleatoria $\mathbf{T}$ tiene distribución exponencial, se escribe $\mathbf{T} \sim \exp(\theta)$.

La función de densidad acumulada es

$$F(t) = \int_0^t f(u)du = 1 - e^{-\theta t}, \ t \geq 0. \tag{3.2.2}$$

La distribución exponencial tiene una propiedad conocida como "falta de memoria", la cual se expresa como

$$P(T \geq t + t_0)|T \geq t) = P(T \geq t_0) = S(t_0).$$

Esto significa que la edad de la unidad (persona, animal u objeto) no afecta la sobrevida futura. Esta característica puede considerarse para casos donde no hay "fatiga" de la unidad.

La función de sobrevida es

$$S(t) = 1 - F(t) = e^{-\theta t}, \ t \geq 0. \tag{3.2.3}$$

Por tanto, la función de riesgo de (1.5.2) es

$$h(t) = \theta, \ t \geq 0. \tag{3.2.4}$$

Se interpreta que el riesgo para la unidad asociada es el mismo o es constante e independiente de la edad de la unidad, característica que es consistente con la falta de memoria de la distribución de esta variable aleatoria (no fatiga o desgaste). La figura 3.1 contiene las gráficas para las tres funciones asociadas con el modelo exponencial. Si se toma el logaritmo natural sobre la función de sobrevida (3.2.3), se tiene que $\ln S(t) = -\theta t$, la cual es una función lineal de t. A partir de esta propiedad resulta sencillo explorar si los datos son generados por un modelo probabilístico exponencial. Basta elaborar una gráfica de $\ln \hat{S}(t)$ frente al tiempo t, donde $\hat{S}(t)$ es un estimador de $S(t)$. Una aproximación a una línea recta de la nube de puntos $(t, \hat{S}(t))$ advierte si los datos siguen una distribución exponencial. La pendiente de la línea recta es un estimador empírico de la tasa de riesgo θ.

La media y la varianza de una distribución exponencial con parámetro θ son $E(\mathbf{T}) = 1/\theta$ y $var(\mathbf{T}) = 1/\theta^2$, respectivamente. La mediana es $1/\theta \ln 2$. El coeficiente de variación es 1, tal como escriben Lee-Wang [44].

3.3. Distribución exponencial generalizada

Una generalización de la distribución exponencial consiste en incluir un parámetro de "garantía". Esto produce la distribución exponencial biparamé-

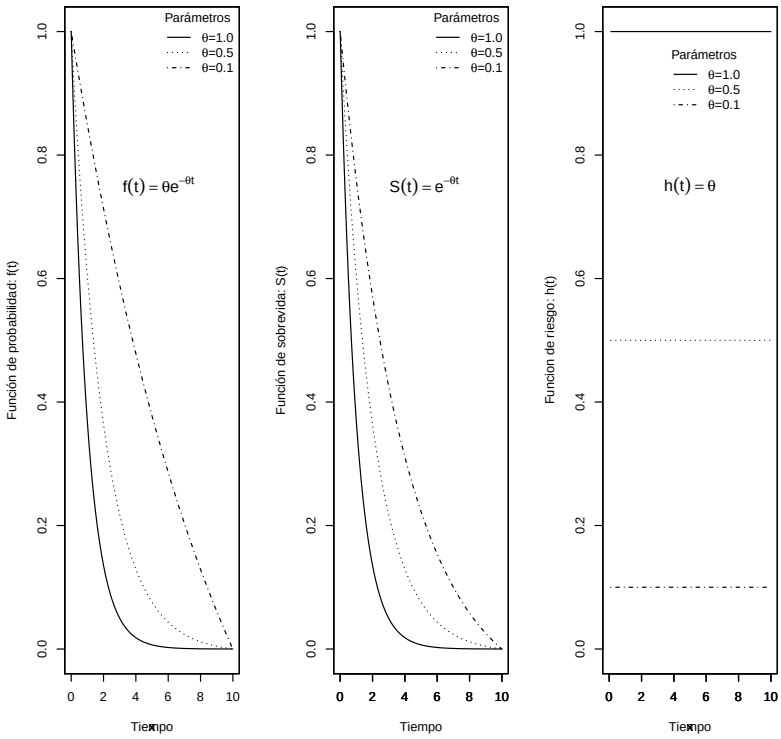

Figura 3.1. Función de probabilidad exponencial, sobrevida y riesgo asociadas.

trica, cuyas funciones de densidad acumulada, sobrevida y riesgo son, respectivamente,

$$f(t) = \begin{cases} \theta e^{-\theta(t-\mathcal{G})}, & t \geq \mathcal{G} \\ 0, & t < \mathcal{G}, \end{cases} \tag{3.3.1}$$

$$F(t) = \begin{cases} 1 - e^{-\theta(t-\mathcal{G})}, & t \geq \mathcal{G} \\ 0, & t < \mathcal{G}, \end{cases} \tag{3.3.2}$$

$$S(t) = \begin{cases} e^{-\theta(t-\mathcal{G})}, & t \geq \mathcal{G} \\ 0, & t < \mathcal{G}, \end{cases} \tag{3.3.3}$$

y

$$h(t) = \begin{cases} 0, & 0 \leq t < \mathcal{G} \\ \theta, & t \geq \mathcal{G}. \end{cases} \tag{3.3.4}$$

El parámetro $\mathcal{G}$ corresponde al *tiempo de garantía*, dentro del cual se asegura que la probabilidad de que el evento ocurra antes del tiempo $\mathcal{G}$ es cero. Para el caso de un dispositivo mecánico o electrónico se afirma que el tiempo de sobrevida del dispositivo es mínimo $\mathcal{G}$. La media y la mediana de esta distribución son $\mathcal{G}+1/\theta$ y $(\ln 2+\theta\mathcal{G})/\theta$, respectivamente. Note que si $\mathcal{G} = 0$, se tienen la media y la varianza de la distribución exponencial.

3.4. Distribución de Weibull

En la década de 1930, las ciencias de la ingeniería introdujeron la distribución de Weibull dentro de los problemas relacionados con valores extremos. En 1939, el ingeniero sueco Waloddi Weibull publicó dos informes sobre la resistencia de materiales. Probablemente, el modelo físico más antiguo que conduce a una distribución de Weibull está conectado con situaciones de valor extremo, y de allí, los orígenes de las distribuciones de valor extremo. Esta distribución se aplica no solo en las ciencias de la salud, sino también en la ingeniería y en áreas como la confiabilidad, la resistencia de materiales y la degradación, entre otras. Los dos libros dedicados exclusivamente a esta distribución son los de Rinne y McCool.

La distribución de Weibull para la variable *tiempo al evento*, **T**, permite tasas de riesgo que pueden ser crecientes, decrecientes o constantes. Es una generalización de la distribución exponencial.

Esta distribución es definida o caracterizada por dos parámetros, θ y α, y se escribe $\mathbf{T} \sim Weibull(\theta, \alpha)$. El valor del parámetro θ determina la escala, mientras que el valor de α determina la forma de la distribución.

La función de densidad de probabilidad es

$$f(t) = \begin{cases} \theta\alpha(\theta t)^{\alpha-1}e^{-(\theta t)^{\alpha}}, & t \geq 0,\ \theta, \alpha > 0 \\ 0, & t < 0. \end{cases} \tag{3.4.1}$$

Note que si $\mathbf{T} \sim Weibull(\theta, \alpha)$, se tienen estas propiedades (intente demostrarlas):

- Si $\alpha = 1$, entonces la variable aleatoria T tiene distribución $\exp(\theta)$.
- La variable aleatoria T^{α} tiene distribución $\exp(\theta^{\alpha})$.

Además, se demuestra que la función de densidad acumulada es

$$F(t) = 1 - e^{-(\theta t)^{\alpha}},\ t \geq 0. \tag{3.4.2}$$

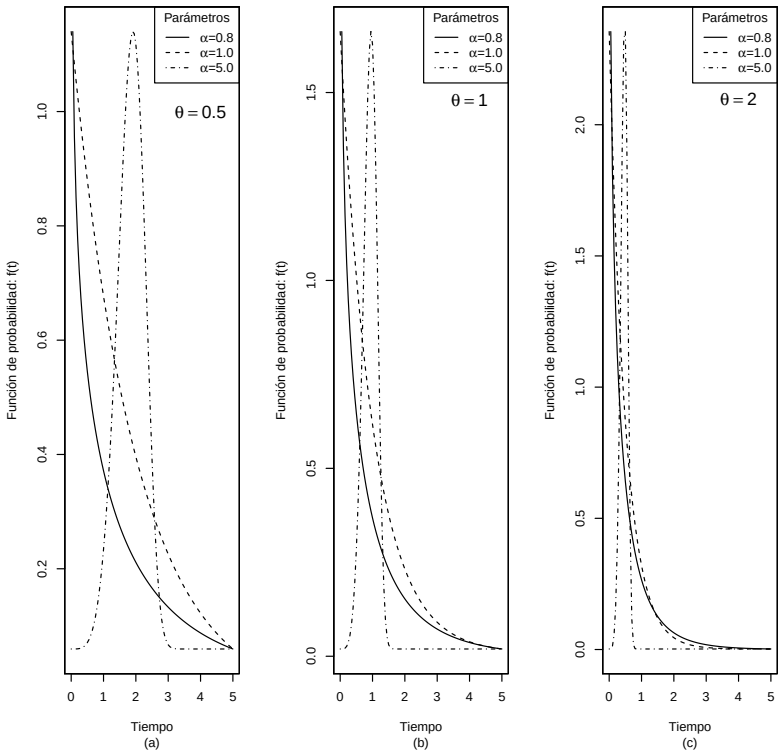

Figura 3.2. Función de probabilidad Weibull.

La función de sobrevida es

$$S(t) = 1 - F(t) = e^{-(\theta t)^{\alpha}}, \ t \geq 0. \tag{3.4.3}$$

Por tanto, la función de riesgo de (1.5.2) es

$$h(t) = \theta\alpha(\theta t)^{\alpha-1}, \ t \geq 0. \tag{3.4.4}$$

Con un poco de cálculo sobre la función $h(t)$ se observa que la función de riesgo Weibull tiene solo alguna de estas características: creciente, decreciente o constante. La derivada de $h(t)$ respecto de t es

$$h'(t) = \theta^2\alpha(\alpha - 1)(\theta t)^{\alpha-2},$$

cuyo signo es positivo si $\alpha > 1$ (pues $t, \theta > 0$), negativo para $\alpha < 1$ y cero para $\alpha = 1$. Así, la función es creciente, decreciente o constante de acuerdo con estos valores de α. La figura 3.4 muestra esta característica para diferentes valores de este parámetro.

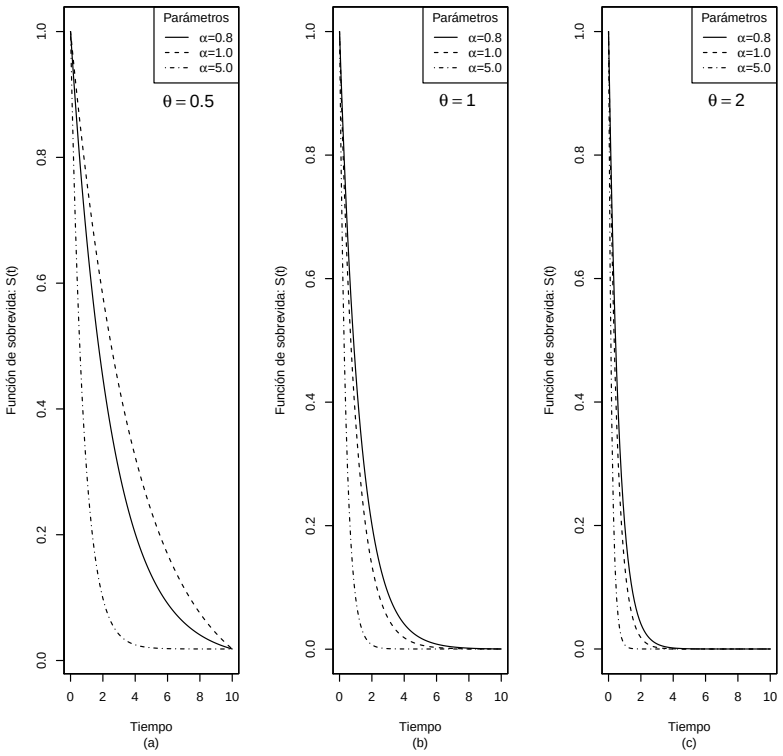

Figura 3.3. Función de sobrevida Weibull.

Mediante la aplicación de logaritmo a la función de sobrevida en (3.4.3), se obtiene

$$\ln(S(t)) = -(\theta t)^{\alpha}.$$

Aplicando nuevamente logaritmos a los dos miembros de la última expresión, se obtiene

$$\ln[-\ln(S(t))] = \alpha \ln \theta + \alpha \ln t.$$

Si el modelo probabilístico que genera los datos es el de Weibull, se espera que la nube de puntos del tipo $(\ln t, \ln[-\ln(S(t))])$ se aproxime a una línea recta, con intercepto $\beta_0 = \alpha \ln \theta$ y pendiente $\beta = \alpha$. Asumiendo que la aproximación de los datos así transformados a una línea recta es razonablemente buena, entonces se pueden obtener estimaciones empíricas de α y θ; estos son

$$\hat{\alpha} = \hat{\beta} \text{ y } \hat{\theta} = e^{\hat{\beta}_0/\hat{\alpha}},$$

donde $\hat{\beta}_0$ y $\hat{\beta}$ son el intercepto y la pendiente de la recta que "mejor" se ajusta a los datos observados.

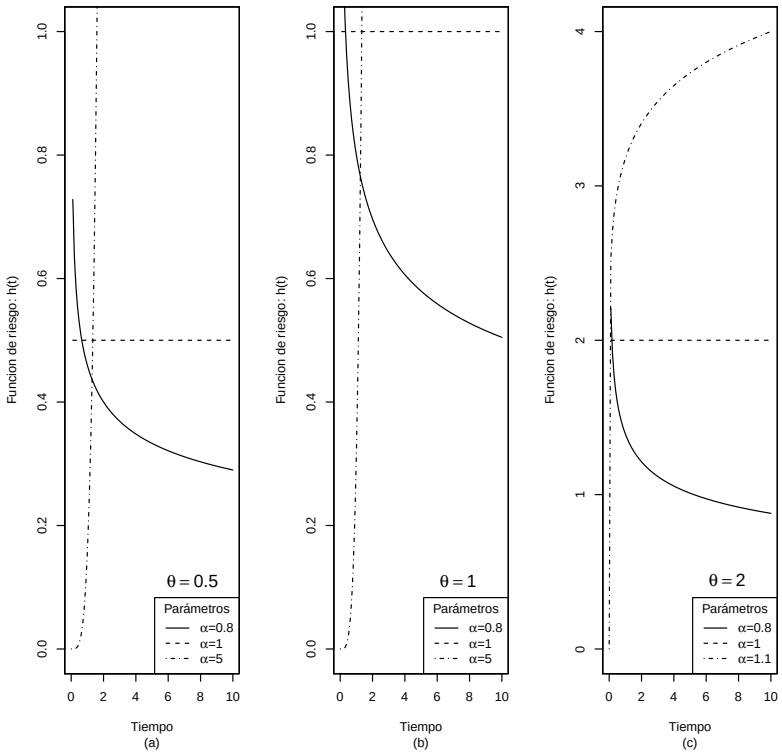

Figura 3.4. Función de riesgo Weibull.

De esta manera, se dispone de un procedimiento heurístico útil para dar cuenta de qué tan apropiado es el modelo probabilístico de Weibull para los datos y de un procedimiento exploratorio de estimación de los parametros.

Igual que se generaliza la distribución exponencial, la distribución de Weibull puede generalizarse a un modelo que considere e incluya un tiempo de "garantía" (en general, un umbral o *threshold*, en inglés); antes de este, se sostiene que la probabilidad de que se presente el evento es cero. Las funciones de densidad, sobrevida y riesgo son, respectivamente,

$$f(t) = \begin{cases} \theta^{\alpha}\alpha(t-\mathcal{G})^{\alpha-1}e^{-\theta^{\alpha}(t-\mathcal{G})^{\alpha}}, & t \geq \mathcal{G} \\ 0, & t < \mathcal{G}, \end{cases} \tag{3.4.5}$$

$$S(t) = \begin{cases} e^{-\theta^{\alpha}(t-\mathcal{G})^{\alpha}}, & t \geq \mathcal{G} \\ 0, & t < \mathcal{G}, \end{cases} \tag{3.4.6}$$

y

$$h(t) = \begin{cases} 0, & 0 \leq t < \mathcal{G} \\ \theta^{\alpha}\alpha(t-\mathcal{G})^{\alpha-1}, & t \geq \mathcal{G}. \end{cases} \tag{3.4.7}$$

La distribución de Weibull se generaliza a: Weibull hasta con cinco parámetros, la inversa Weibull, log-Weibull, exponencial Weibull, beta Weibull, Weibull-Poisson, entre otras. La monografía de Lai [40] presenta un número considerable de extensiones de la distribución Weibull donde se incluyen las citadas aquí.

3.5. Distribución de log-normal

Así como la distribución de Weibull, la distribución log-normal es muy empleada para caracterizar tiempos de vida de unidades, artefactos o individuos, tales como el tiempo hasta la fatiga de un metal o de un componente mecánico o electrónico, o el tiempo de vida de pacientes con una enfermedad.

La distribución de una variable aleatoria T es log-normal cuando la distribución de su logaritmo, $\mathbf{Y} = \ln(\mathbf{T})$, tiene una distribución normal.

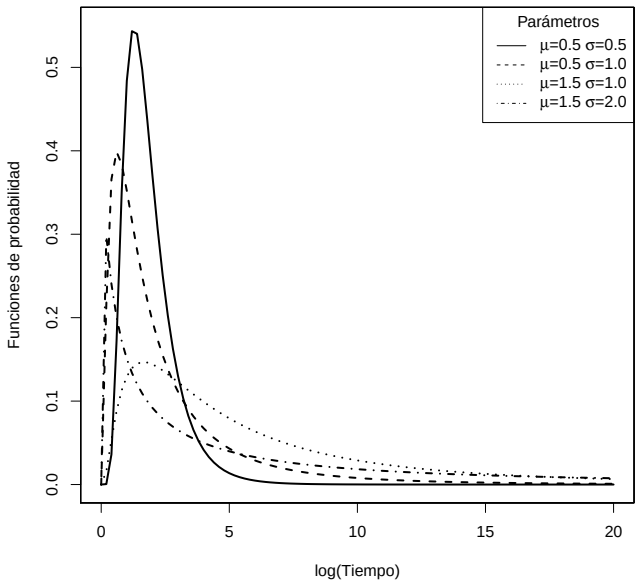

Figura 3.5. Función de probabilidad log-normal.

Su función de densidad queda completamente especificada por los parámetros μ y σ, los cuales corresponden a la media y la varianza de **Y**, y está dada por:

$$f(t) = \frac{1}{t\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{\ln(t)-\mu}{\sigma}\right)^2\right] \tag{3.5.1}$$

$$= \frac{\phi\left(\frac{\ln(t)-\mu}{\sigma}\right)}{t}, \tag{3.5.2}$$

donde $-\infty < \mu < \infty$, $0 < \sigma < \infty$ y $\phi(\cdot)$ es la función de densidad de una variable aleatoria con distribución normal estándar.

La función de sobrevida está dada por:

$$S(t) = 1 - \Phi\left(\frac{\ln(t)-\mu}{\sigma}\right) = \Phi\left(\frac{-\ln(t)+\mu}{\sigma}\right), \tag{3.5.3}$$

donde Φ es la función de distribución acumulada de una variable aleatoria con distribución normal estándar.

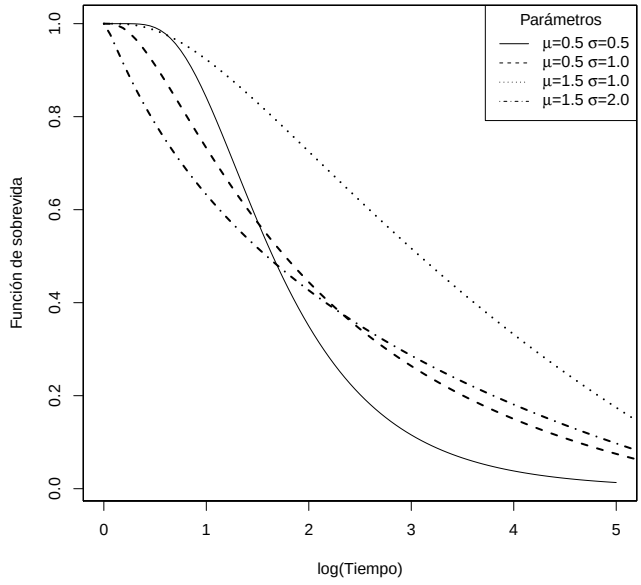

Figura 3.6. Función de sobrevida log-normal.

La función de riesgo de la distribución log-normal es

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{\phi\left(\frac{\ln(t)-\mu}{\sigma}\right)}{t}}{1 - \Phi\left(\frac{\ln(t)-\mu}{\sigma}\right)},$$

la cual tiene forma de "montaña", dado que toma el valor cero al tiempo cero, después crece hasta alcanzar un máximo y decrece a cero cuando t tiende al infinito. Esta distribución ha sido criticada para modelar algunos tiempos de falla, dado que la función de riesgo es decreciente para valores grandes de t, lo cual es inaceptable en muchas situaciones. El modelo puede ser factible cuando valores grandes de tiempo no son de interés.

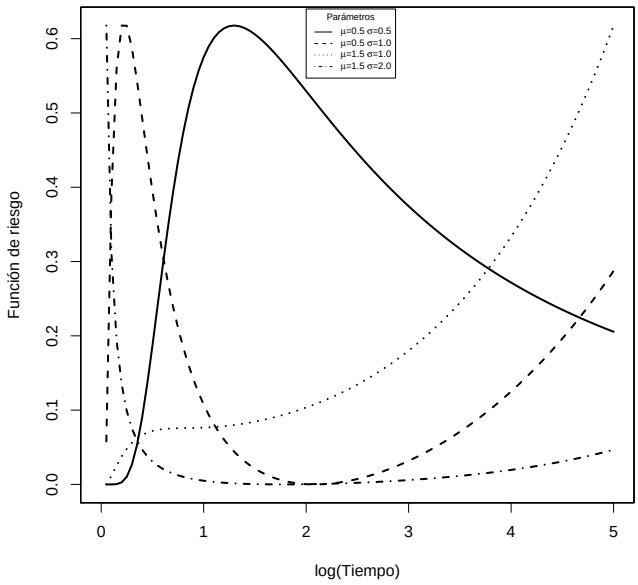

Figura 3.7. Función de riesgo log-normal.

3.6. Distribución log-logística

La función de riesgo de Weibull se caracteriza por ser creciente, decreciente (monótona) o constante, como se observa en la figura 3.4. No obstante, se presentan situaciones donde el riesgo cambia su sentido, de creciente a decreciente, o recíprocamente, de decreciente a creciente. Por ejemplo, en

ciertas cirugías sobre humanos, el riesgo de muerte se incrementa durante los primeros días del postoperatorio, después de cierto tiempo (una vez que el organismo se recupera) el riesgo de muerte disminuye. Otro ejemplo puede ser el efecto de un fármaco suministrado a un paciente, pues durante el periodo inicial de su ingesta puede tener el efecto esperado en la salud (baja el riesgo de muerte), pero luego este efecto puede disminuir e incluso volverse inhibitorio (aumenta el riesgo de muerte). Una situación recíproca se presenta, en el caso de un estudiante universitario, con el evento "deserción académica". Cuando el estudiante inicia sus estudios en la universidad, el riesgo de deserción puede ir en aumento, después se presenta una adaptación al sistema universitario y este riesgo disminuye gradualmente (es decir, obtiene el grado). En estos casos se requiere de una función de riesgo que tenga una forma unimodal.

Así, la importancia de la distribución log-logística, como la log-normal, está en que proporciona modelos para tiempos de vida que no tienen función de riesgo monótona, sino que pueden (de acuerdo a como se elijan los parámetros) presentar diferentes formas. Las funciones de riesgo de estas dos distribuciones son muy parecidas, pero la de log-logística es mucho más manejable.

La variable aleatoria $\mathbf{T}$ sigue una distribución *log-logística* si su logaritmo, $\mathbf{W} = \ln \mathbf{T}$, sigue una distribución logística (una distribución similar a la normal).

La función de densidad de probabilidad de una variable aleatoria con distribución log-logística es

$$f(t) = \frac{\alpha \gamma t^{\gamma-1}}{(1+\alpha t^{\gamma})^2}, \text{ con } t, \alpha \text{ y } \gamma > 0. \tag{3.6.1}$$

En algunas ocasiones es conveniente trabajar con el logaritmo de los tiempos observados, pues se reduce la escala. Se tiene así que la variable aleatoria $\mathbf{T}$ se transforma conforme a $\mathbf{W} = \ln(\mathbf{T})$, la cual tiene una distribución *log-logística* con parámetros α y $\gamma > 0$. La función de densidad de probabilidad de $\mathbf{W}$ es

$$f(w) = \frac{\exp\left(\frac{w-\mu}{\sigma}\right)}{\sigma\left[1+\exp\left(\frac{w-\mu}{\sigma}\right)\right]^2}, \text{ para } -\infty < w < \infty, \tag{3.6.2}$$

con $-\infty < \mu < \infty$ y $\sigma > 0$, los respectivos parámetros de localización y escala.

La figura 3.8 muestra cuatro funciones de densidad de probabilidad de una variable aleatoria log-logística definida para $\alpha = 25$ y $\gamma = 4, 3, 2, 1,$

respectivamente. Se observa que para valores grandes de γ la forma de la distribución tiende a una curva simétrica.

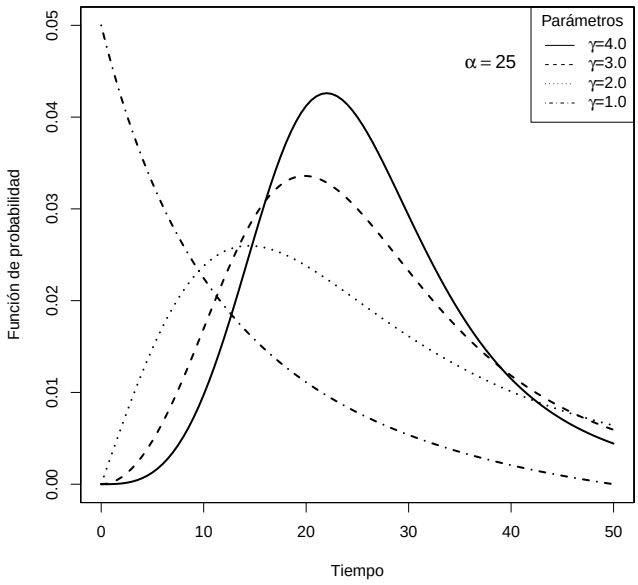

Figura 3.8. Función de densidad de probabilidad log-logística.

La función de sobrevida o de confiabilidad es

$$S(t) = \frac{1}{1 + \left(\frac{t}{\alpha}\right)^{\gamma}}. \tag{3.6.3}$$

La expresión equivalente para la transformación $\mathbf{W} = \ln(\mathbf{T})$ es

$$S(w) = \frac{1}{1 + \exp\left(\frac{w-\mu}{\sigma}\right)}. \tag{3.6.4}$$

La figura 3.9 contiene las funciones de sobrevida para el párametro de localización $\alpha = 25$ y los parámetros de forma $\gamma = 4, 3, 2, 1$. La función de riesgo o tasa de riesgo para el evento de una variable aleatoria log-logística se calcula mediante la expresión

$$h(t) = \frac{\gamma(t/\alpha)^{\gamma-1}}{\alpha\left[1 + (t/\alpha)^{\gamma}\right]}. \tag{3.6.5}$$

Para la transformación $\mathbf{W} = \ln(\mathbf{T})$, la función de riesgo equivalente a la anterior es

$$h(t) = \frac{1}{\sigma} \exp\left\{\frac{w-\mu}{\sigma}\right\} \left(1 + exp\left\{\frac{w-\mu}{\sigma}\right\}\right)^{-1}. \tag{3.6.6}$$

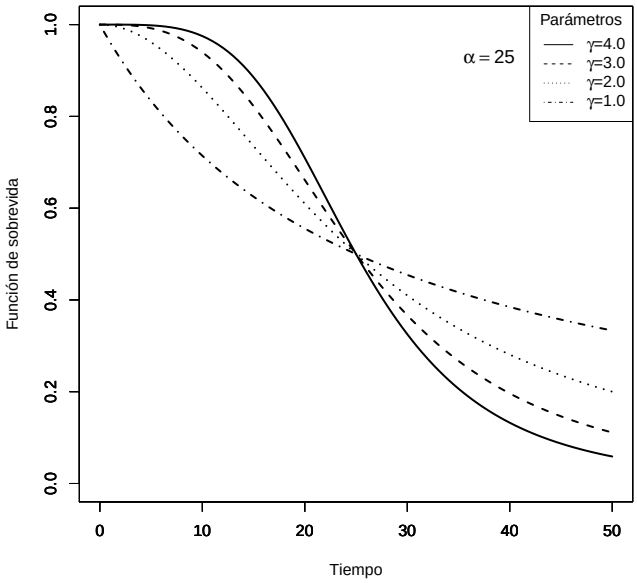

Figura 3.9. Función de sobrevida log-logística.

Los parámetros asociados en función de t y en función de w (log-logística y logística) se relacionan como $\gamma = 1/\sigma$ y $\alpha = \exp(\mu)$.

Esta función de riesgo tiene una similitud con la correspondiente a una variable aleatoria log-normal, ya que, inicialmente, el riesgo crece hasta alcanzar un punto máximo y luego decrece.

Esta función decrece monotónicamente si $\gamma < 1$, mientras que si $\gamma > 1$ la función de riesgo tiene una única moda. Así, la función de riesgo log-logística se puede emplear en situaciones donde primero el riesgo crece hasta un pico (punto de riesgo máximo) y luego decrece, o también en situaciones donde el riesgo es monotónicamente decreciente.

La figura 3.10 corresponde a la funciones de riesgo para los parámetros $\alpha = 25$ y $\gamma = 0.5, 1, 3, 4$, respectivamente.

El p-ésimo percentil o el cuantil de orden p se calcula mediante

$$\xi_p = \alpha \left[\frac{p}{1-p} \right]^{1/\gamma}. \tag{3.6.7}$$

La mediana corresponde a $p = 0.5$. De la igualdad (3.6.7), el cuantil asociado con la mediana es $\xi_{0.5} = \alpha$.

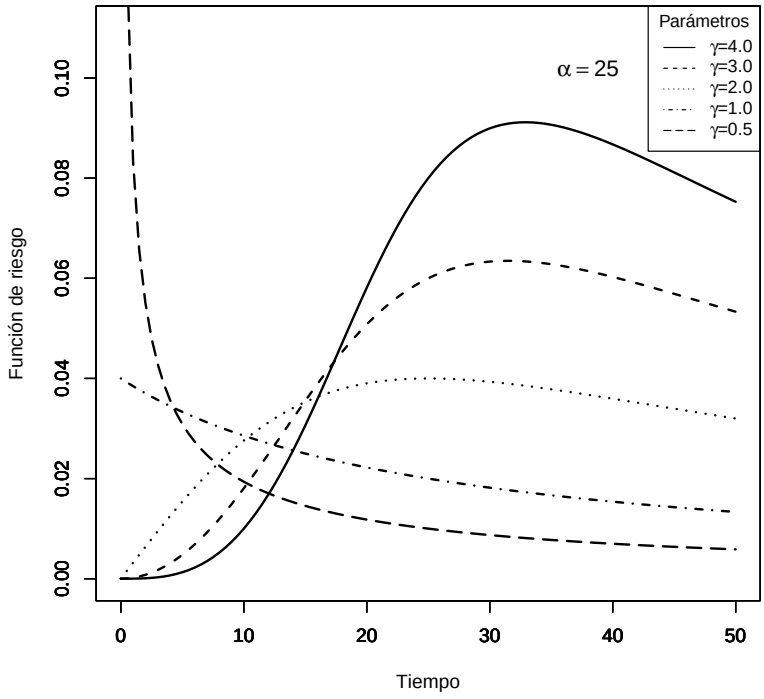

Figura 3.10. Función de riesgo log-logística.

3.7. Distribución gama

Como se ha referido anteriormente, la distribución de Weibull es una generalización biparamétrica de la distribución exponencial. Así mismo, otra generalización es la *distribución gama*. Esta distribución ha sido empleada para describir el tiempo de vida de componentes electrónicos, es decir, se usa en problemas de confiabilidad. No obstante, más recientemente ha sido empleada en el área de la medicina para modelar el tiempo de vida de pacientes. La función de densidad es

$$f(t) = \frac{\theta^{\alpha} t^{\alpha-1} e^{-\theta t}}{\Gamma(\alpha)}, \text{ con } t, \theta, \alpha > 0. \tag{3.7.1}$$

Se escribe $\mathbf{T} \sim \mathcal{G}(\theta, \alpha)$ para denotar que $\mathbf{T}$ tiene una distribución gama con parámetros θ y α; estos son, respectivamente, parámetros de escala y forma.

Tal como en la distribución de Weibull, θ es el parámetro de localización y α es el parámetro de escala. Esta distribución, también como la de Weibull, contiene la distribución exponencial como un caso especial si $\alpha = 1$. Además, la distribución gama se aproxima a la distribución normal si $\alpha \to \infty$ y es una distribución ji-cuadrado con ν grados de libertad cuando $\nu = 2\alpha$ y $\theta = 1/2$. La media y la varianza de la distribución gama son α/θ y α/θ^2, respectivamente.

Se debe distinguir entre la función de densidad gama y la función gama. La primera es definida para una variable aleatoria **T** en los números reales positivos, esta es dada por (3.7.1), mientras que la función gama, $\Gamma(\cdot)$, se define por

$$\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1}e^{-t}dt, \text{ para } t, \alpha > 0. \tag{3.7.2}$$

La función gama tiene amplias aplicaciones en el campo matemático y en particular dentro de la teoría de probabilidades. Las propiedades que se enuncian a continuación pueden ser demostradas sin mayor complejidad. La propiedad i se obtiene por la evaluación de la integral, para la propiedad ii se sugiere hacer $I = \int_0^{+\infty} t^{\alpha-1}e^{-t}dt$, luego desarrollar I^2 (transformando por coordenadas polares) y obtener nuevamente I. Para la propiedad iii se debe hacer integración por partes y, finalmente, la propiedad iv se obtiene de iii haciendo $\alpha = n$, un entero no negativo.

i) $\Gamma(1) = 1$,

ii) $\Gamma(\frac{1}{2}) = \sqrt{\pi}$,

iii) $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$, para α cualquier número real positivo,

iv) $\Gamma(n + 1) = n!$, para $n = 0, 1, 2, \ldots$, un entero no negativo.

La figura 3.11 contiene diferentes gráficas de funciones de densidad de probabilidad tipo gama para el parámetro de localización, $\theta = 1$, y diferentes valores del parámetro de escala $\alpha = 0.5, 1.0, 1.5, 3.5$, respectivamente. Como se ha señalado arriba, para $\alpha = 1$ se tiene la distribución exponencial y para valores de α grandes la distribución converge asintóticamente a la normal; aunque un valor de $\alpha = 3.5$ no es muy grande, por lo menos visualmente se insinúa un acercamiento a la curva de una distribución normal.

Para valores de $\alpha > 1$ hay un punto máximo en $t = (\alpha - 1)/\theta$. En este punto "pico" o crítico se debe considerar que para valores de $\alpha < 1$ se tendrían tiempos o edades con valor negativo.

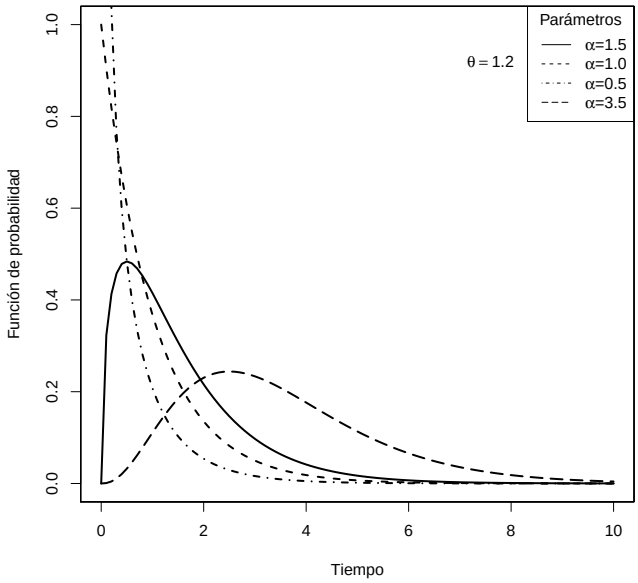

Figura 3.11. Función de densidad de probabilidad gama.

La función de sobrevida o de confiabilidad de la variable aleatoria **T** con distribución gama es

$$\begin{aligned} S(t) &= \left[\int_t^{+\infty} \theta(\theta t)^{\alpha-1} \exp(-\theta t) dt\right] / \Gamma(\alpha) \\ &= 1 - \left[\int_0^{\theta x} u^{\alpha-1} \exp(-u) du\right] / \Gamma(\alpha) \\ &= 1 - I(\theta x, \alpha), \end{aligned} \tag{3.7.3}$$

donde, $u = \theta t$ e $I(\theta x, \alpha)$ se denomina la *función gama incompleta*.

La figura 3.12 exhibe diferentes funciones de sobrevida de acuerdo con los valores del parámetro de localización $\theta = 1$, y los valores del parámetro de escala $\alpha = 0.5, 1.0, 1.5, 3.5$, respectivamente. Se puede advertir que el evento ocurre de manera relativamente prematura para valores de α pequeños; es decir que los valores grandes de α están asociados con la ocurrencia del evento progresiva y probablemente "tardía". Para valores de γ enteros ($\gamma = n$), se obtiene la *distribución de Erlang*, cuyas funciones de sobrevida y

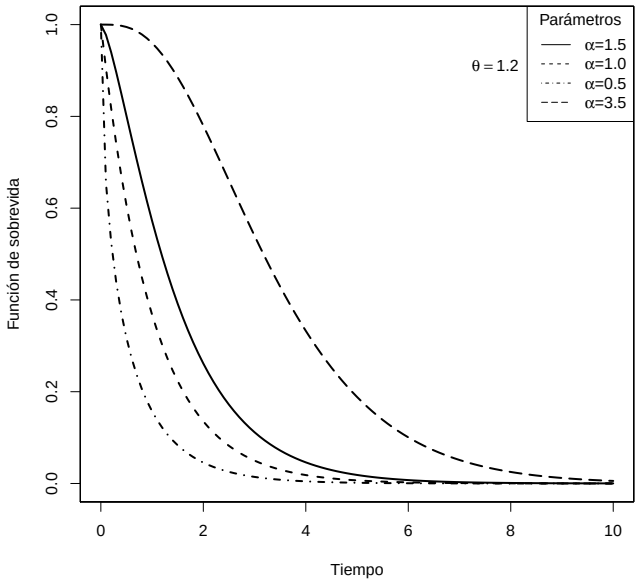

Figura 3.12. Función de sobrevida gama.

riesgo, respectivamente, son

$$S(t) = \exp(\theta t) \sum_{k=0}^{n-1} (\theta t)^k / k!, \tag{3.7.4}$$

y

$$h(t) = \theta(\theta t)^{n-1} \left[(n-1) \sum_{k=0}^{n-1} (\theta t)^k / k! \right]^{-1}. \tag{3.7.5}$$

La función de riesgo para la variable aleatoria, **T**, con distribución gama se obtiene de la identidad $h(t) = f(t)/S(t)$.

La gráfica 3.13 muestra las funciones de riesgo para el parámetro de localización $\theta = 1$ y los mismos valores anteriores del parámetro de escala $\alpha = 0.5, 1.0, 1.5, 3.5$. Desde la gráfica se puede corroborar que para valores del parámetro de escala, α, menores que 1.0, el riesgo decae progresivamente con el tiempo hasta un valor mínimo del parámetro de localización θ. Para el valor de $\alpha = 1$ la tasa de riesgo se hace constante e igual a θ (el caso de la distribución exponencial). Cuando $\alpha > 1$, el riesgo se aumenta monotónicamente desde 0 hasta θ. En resumen, la distribución gama describe diferentes patrones de sobrevida asociados a tasas de riesgo que pueden ser

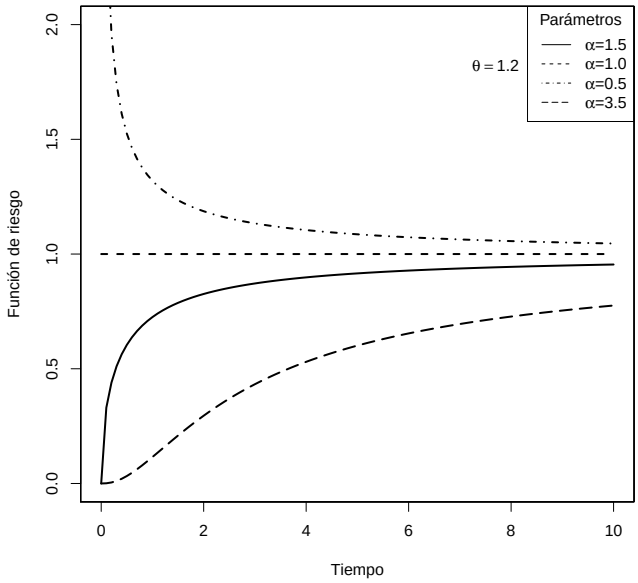

Figura 3.13. Función de riesgo gama.

decrecientes o crecientes hasta alcanzar un valor constante a medida que el tiempo transcurre.

3.8. Distribución gama generalizada

Es una distribución triparamétrica, con parámetros θ, α y γ, cuya función de densidad de probabilidad es definida como

$$f(t) = \frac{\gamma\theta^{\alpha}}{\Gamma(\alpha)} t^{\gamma\alpha-1} \exp[-\theta t^{\gamma}], \text{ con } t, \theta, \alpha, \gamma > 0. \tag{3.8.1}$$

Las distribuciones exponencial, Weibull, log-normal y gama son un caso especial de esta distribución. Así, la distribución gama generalizada es la distribución exponencial si $\alpha = \gamma = 1$, es la distribución Weibull si $\alpha = 1$, es la distribución log-normal cuando $\alpha \to \infty$, y si $\gamma = 1$, es la distribución gama.

3.9. Distribución de valor extremo

Una variable aleatoria **T** tiene un modelo de probabilidad de *valor extremo* si la forma de su función de riesgo es

$$h(t) = \frac{1}{\sigma} \exp\left(\frac{t-\mu}{\sigma}\right), \tag{3.9.1}$$

para $-\infty < t < \infty$ y los parámetros $\sigma > 0$ y $-\infty < \mu < \infty$.

Se escribe $\mathbf{T} \sim VE(\mu, \sigma)$ para indicar que la variable aleatoria **T** tiene distribución de valor extremo (abreviado como VE) con μ parámetro de localización y σ parámetro de escala.

Note que la variable aleatoria tiempo, **T**, toma valores en todos los numéros reales; es decir, se admiten tiempos negativos.

Con esta función de riesgo se obtiene la función de sobrevida mediante la igualdad (1.5.4), esta es

$$S_1(t) = exp\left\{-\int_{-\infty}^{t} h(u)du\right\} = \exp\left[-\exp\left(\frac{t-\mu}{\sigma}\right)\right], \ -\infty < t < \infty.$$

Se ha advertido acerca del dominio de la distribución de valor extremo (todos los números reales), de forma tal que debe hacerse un truncamiento que redefina la función de sobrevida para valores del tiempo positivos (es decir, $t > 0$). De esta manera, la función de sobrevida es

$$S_2(t) = \exp\left[-\exp\left(\frac{t-\mu}{\sigma}\right) + \exp\left(\frac{-\mu}{\sigma}\right)\right], \ t > 0.$$

El sumando extra en el exponente es una constante de normalización, la cual garantiza que el área bajo la función de densidad de probabilidad sea igual a 1.

Si se expresa

$$\mathbf{Z} = \frac{\mathbf{T}-\mu}{\sigma}, \text{ entonces es equivalente a } \mathbf{T} = \mu + \sigma\mathbf{Z}.$$

De esta manera, si $\mathbf{T} \sim VE(\mu, \sigma)$, entonces $\mathbf{Z} \sim VE(0, 1)$. A **Z** se le denomina modelo de probabilidad de valor extremo estandarizado.

Una propiedad importante de la distribución de Weibull es su transformación logarítmica. Así, suponga que la variable aleatoria *tiempo para un evento*, $\mathbf{T} \sim Weibull(\theta, \alpha)$. Se obtienen la función de sobrevida $S(y)$, con

$Y = \ln \mathbf{T}$, de esta manera:

$$\begin{aligned} S(y) &= P(Y > y) \\ &= P(\ln \mathbf{T} > y) \\ &= P(\mathbf{T} > e^{y}) \\ &= \exp\{-\left(e^{\theta y}\right)^{\alpha}\} \\ &= \exp\{-(\theta)^{\alpha} e^{\alpha y}\} \\ &= \exp[-\exp(\alpha y + \alpha \ln \theta)] \\ &= \exp[-\exp\left(\frac{y-u}{b}\right)], \end{aligned}$$

donde $b = \frac{1}{\alpha}$ y $u = -\ln \theta$. Con esta reparametrización, se tiene que $b > 0$ y $-\infty < u < \infty$. La función de densidad de probabilidad de Y es dada por

$$\begin{aligned} f(y) &= -S'(y) \\ &= \frac{1}{b} \exp\left[\frac{y-u}{b} - \exp\left(\frac{y-u}{b}\right)\right] \\ &= \frac{1}{b} \exp\left(\frac{y-u}{b}\right) S(y), \end{aligned}$$

con $-\infty < u < \infty$.

En conclusión, si $\mathbf{T} \sim Weibull(\theta, \alpha)$, entonces $Y = \ln \mathbf{T}$ tiene una distribución de valor extremo.

3.10. Distribución de Birnbaum-Saunders

La distribución de *Birnbaum-Saunders* fue motivada tanto por la observación de la fatiga de los materiales presentes en las aeronaves comerciales causada por el sometimiento a vibraciones constantes, como por los problemas provocados por esta.

Este modelo ha recibido bastante atención durante los últimos años debido a su capacidad para incluir, de manera eficiente, las características asociadas con fenómenos en los que una fuerza acumulativa excede un umbral (en inglés, *threshold*) de resistencia crítico, algo muy común en campos como la ingeniería, la medicina y el análisis ambiental.

Birnbaum y Saunders derivaron, en 1968, un modelo probabilístico que describe los tiempos de vida asociados con muestras de materiales expuestos a fatiga o degradación, como resultado del trabajo cíclico y la tensión. El año siguiente, en 1969, formalizaron la distribución de vida con aplicación a la fatiga de materiales que más tarde tomó sus nombres. La caracterizaron

como una distribución y propusieron un método de estimación para los parámetros de esta distribución biparamétrica, como se reseña en el libro de Leyva [45, p. 4].

El modelo probabilístico de *Birnbaum-Saunders* (en adelante, BS) corresponde a la distribución de probabilidad de la variable aleatoria:

$$\mathbf{T} = \beta\left(\frac{\alpha}{2}Z + \sqrt{\left(\frac{\alpha}{2}Z\right)^2 + 1}\right)^2, \ \alpha, \beta, t > 0, \tag{3.10.1}$$

donde $Z \sim n(0, 1)$. Se escribe $\mathbf{T} \sim BS(\alpha, \beta)$, donde $\alpha > 0$ representa el parámetro de forma y $\beta > 0$ es tanto el parámetro de escala como la mediana de la distribución. De la expresión (3.10.1),

$$Z = \frac{1}{\alpha}\left(\sqrt{\frac{\mathbf{T}}{\beta}} - \sqrt{\frac{\beta}{\mathbf{T}}}\right) \sim n(0, 1). \tag{3.10.2}$$

La figura 3.14 muestra diferentes gráficas de las funciones de densidad para valores del parámetro de forma $\alpha = 0.1,\ 0.2,\ 0.5,\ 1,\ 2$ y 5, y del parámetro de escala $\beta = 1$. Se advierte que para β fijo y α pequeño la distribución tiende a ser simétrica, con β como el punto de simetría.

Si la variable aleatoria $\mathbf{T} \sim BS(\alpha, \beta)$, entonces su función de distribución acumulada de probabilidad es:

$$F(t) = \Phi\left(\frac{1}{\alpha}\left[\sqrt{\frac{t}{\beta}} - \sqrt{\frac{\beta}{t}}\right]\right), \ \text{con } t, \alpha, \beta > 0, \tag{3.10.3}$$

donde la función $\Phi(\cdot)$ es la función de distribución acumulada de una variable aleatoria Z con distribución normal estándar. La función de densidad de probabilidad de $\mathbf{T}$ se obtiene derivando (3.10.3) respecto a t, esto es

$$f(t) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2\alpha^2}\left[\frac{t}{\beta} + \frac{\beta}{t} - 2\right]\right)\frac{t^{-3/2}(t+\beta)}{2\alpha\sqrt{\beta}}, \ \text{con } t, \alpha, \beta > 0. \tag{3.10.4}$$

Una forma equivalente es

$$f(t) = \phi(a_t)A_t, \text{ donde } A_t = \frac{da_t}{dt} = \frac{t^{-3/2}(t+\beta)}{2\alpha\sqrt{\beta}},$$

es decir, A_t es la derivada interna de $\phi(\cdot)$. Si $\mathbf{T} \sim BS(\alpha, \beta)$, entonces la media y la varianza de $\mathbf{T}$ son, respectivamente,

$$E(\mathbf{T}) = \beta\left(1 + \frac{\alpha^2}{2}\right), \ \text{y } var(\mathbf{T}) = \beta^2\alpha^2\left(\frac{5}{4}\alpha^2 + 1\right).$$

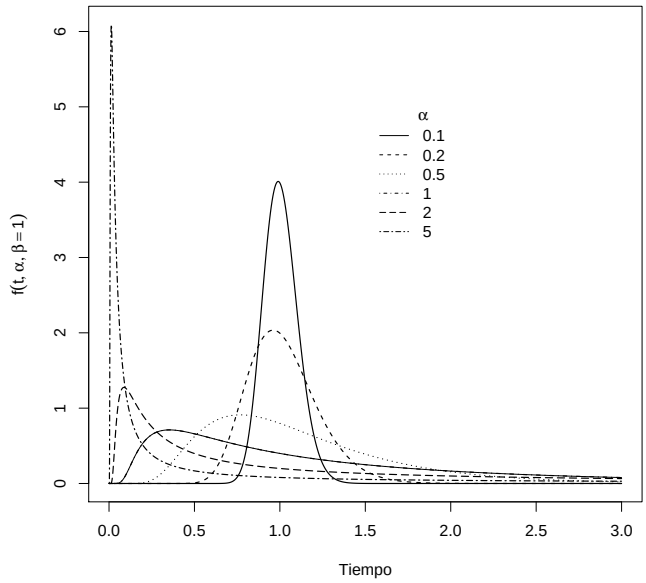

Figura 3.14. Funciones de densidad Birnbaum-Saunders.

La función de sobrevida o de confiabilidad se construye desde la función de distribución acumulada, $F(t)$, definida en (3.10.3), esta es

$$S(t) = 1 - F(t) = \Phi\left(-\frac{1}{\alpha}\left[\sqrt{\frac{t}{\beta}} - \sqrt{\frac{\beta}{t}}\right]\right). \tag{3.10.5}$$

La función de riesgo se construye mediante la expresión (1.5.2), es decir,

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2\alpha^2}\left[\frac{t}{\beta} + \frac{\beta}{t} - 2\right]\right)\frac{t^{-3/2}(t+\beta)}{2\alpha\sqrt{\beta}}}{\Phi\left(-\frac{1}{\alpha}\left[\sqrt{\frac{t}{\beta}} - \sqrt{\frac{\beta}{t}}\right]\right)}. \tag{3.10.6}$$

Esta función tiene estas tres características:

- La función de riesgo, $h(t)$, es unimodal para cualquier valor del parámetro de escala α.
- $h(t)$ crece (aumenta el punto "pico" o modal) en tanto α sea grande.
- $h(t)$ se aproxima al valor $1/(2\alpha^2\beta)$ cuando $t \to \infty$.

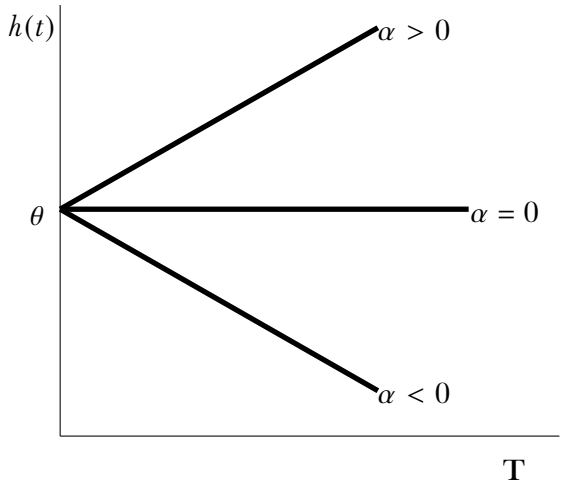

Figura 3.15. Función de riesgo lineal-exponencial.

3.11. Otras distribuciones de sobrevida

Las distribuciones que se presentan en esta sección se construyen desde la función de riesgo, $h(t)$. Tanto la función de sobrevida, $S(t)$, como la función de probabilidad, $f(t)$, asociadas con $h(t)$, se obtienen de las igualdades (1.5.4) y (1.5.2). Algunas distribuciones adicionales se pueden consultar en Cox y Oakes [16].

La función de riesgo de la distribución *lineal-exponencial* es

$$h(t) = \theta + \alpha t, \tag{3.11.1}$$

donde θ y α deben ser valores tales que hagan que $h(t)$ sea no negativa.

La tasa de riesgo crece desde θ si $\alpha > 0$, decrece desde θ si $\alpha < 0$, permanece constante (caso exponencial) si $\alpha = 0$. La figura 3.15 muestra esta caracterización de $h(t)$.

El modelo lineal-exponencial es una extensión del modelo exponencial. El patrón de la sobrevida o del tiempo para el evento tiene un riesgo inicial constante y la tasa de riesgo varía como una función lineal en el tiempo o la edad.

La función de sobrevida se obtiene de la igualdad (1.5.4), esta es

$$\begin{aligned} S(t) &= e^{-\int_o^t h(u)du} \\ &= e^{-\int_o^t (\theta+\alpha u)du} \\ &= e^{[-(\theta t+\frac{1}{2}\alpha t^2)]}. \end{aligned} \tag{3.11.2}$$

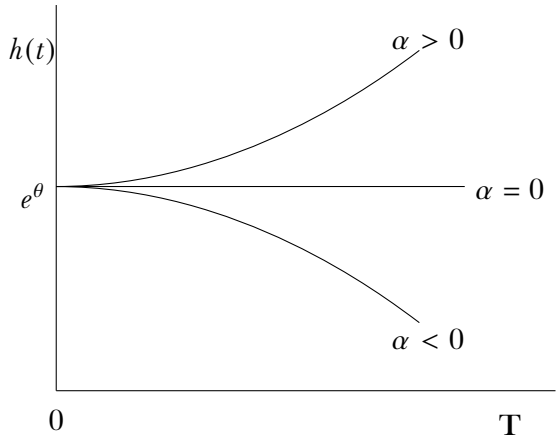

Figura 3.16. Función de riesgo Gompertz.

De acuerdo con la igualdad (1.5.2), $f(t) = h(t)S(t)$, por lo que la función de densidad de probabilidad es

$$f(t) = (\theta + \alpha t) \exp\left[-(\theta t + \frac{1}{2}\alpha t^2)\right]. \tag{3.11.3}$$

Un caso especial de la distribución lineal-exponencial es la distribución de *Rayleigh*, la cual se obtiene reemplazando α por $\frac{1}{2}\alpha$, de manera que la función de riesgo de la distribución de Rayleigh es $h(t) = \theta + \frac{1}{2}\alpha t$. Un procedimiento similar al seguido para el modelo lineal-exponencial permite obtener la función de sobrevida y la función de densidad de probabilidad de Rayleigh.

La distribución de *Gompertz* tiene asociada una función de riesgo cuya forma funcional es

$$h(t) = \exp(\theta + \alpha t). \tag{3.11.4}$$

Los valores de θ y α deben ser tales que $h(t)$ sea no negativa. La tasa de riesgo varía como función exponencial del tiempo o la edad de una unidad. Tanto la función de riesgo como la de sobrevida, ligadas a la distribución de Gompertz, son una extensión de las análogas para la exponencial.

La figura 3.16 muestra la función de riesgo tipo Gompertz. Para $\alpha > 0$ hay edades o tiempos positivos y la tasa de riesgo inicia en e^{θ}; cuando $\alpha < 0$, existen edades negativas y a la edad $t = 0$ el riesgo es igual a e^{θ}. Para el valor $\alpha = 0$, la función de riesgo, $h(t)$, se mantiene constante en e^{θ}.

Desde la igualdad (1.5.4), la función de sobrevida o confiabilidad Gompertz es

$$S(t) = \exp\left[-\frac{e^{\theta}}{\alpha}(e^{\theta t} - 1)\right]. \tag{3.11.5}$$

También, desde (1.5.2) y (3.11.4) la función de densidad de probabilidad Gompertz es

$$f(t) = \exp\left[(\theta + \alpha t) - \frac{1}{\alpha}(e^{\theta+\alpha t} - e^{\theta})\right]. \qquad (3.11.6)$$

Tabla 3.1. Algunos modelos probabilísticos para el A. S.

Distribución	Función de riesgo, $h(t)$	Función de sobrevida, $S(t)$	Función de densidad, $f(t)$
Exponencial	θ	$\exp(-\theta t)$	$\theta \exp(-\theta t)$
Weibull	$\alpha\theta t^{\alpha-1}$	$\exp(-\theta t^{\alpha})$	$\alpha\theta t^{\alpha-1}\exp(-\theta t^{\alpha})$
Gamma	$\frac{f(t)}{S(t)}$	$1 - I(\theta t, \alpha)$	$\frac{\theta^{\alpha} t^{\alpha-1}\exp^{(-\theta t)}}{\Gamma(\alpha)}$
Log–normal	$\frac{f(t)}{S(t)}$	$1\text{-}\Phi\left(\frac{\ln(t)-\mu}{\sigma}\right)$	$\frac{\phi\left(\frac{\ln(t)-\mu}{\sigma}\right)}{t}$
Log–logistic	$\frac{\gamma\alpha t^{\gamma-1}}{1+\alpha t^{\gamma}}$	$\frac{1}{1+\alpha t^{\gamma}}$	$\frac{\gamma\alpha t^{\gamma-1}}{[1+\alpha t^{\gamma}]^2}$
Gompertz	$\theta e^{\alpha t}$	$\exp\left[\frac{\theta}{\alpha}(1 - e^{\alpha t})\right]$	$\theta e^{\alpha t}\exp\left[\frac{\theta}{\alpha}(1 - e^{\alpha t})\right]$
Pareto	$\frac{\theta}{t}$	$\frac{\lambda^{\theta}}{t^{\theta}}$	$\frac{\theta\lambda^{\theta}}{t^{\theta+1}}$
Gama-Gen.	$\frac{f(t)}{S(t)}$	$1\text{-}I\,(\theta t^{\gamma}, \alpha)$	$\frac{\gamma\theta^{\alpha} t^{\gamma\alpha-1}\exp\,(-\theta t^{\gamma})}{\Gamma(\alpha)}$
donde $I(t, \alpha) = \int_0^t u^{\alpha-1}\exp(-u)du/\Gamma(\alpha)$			

En la tabla 3.1 se resumen algunas de las distribuciones de uso más frecuente en el análisis de sobrevida.

Capítulo

cuatro

Estimación de parámetros en modelos de sobrevida

$$h(t)=\frac{f(t)}{S(t)}=\frac{f(t)}{1-F(t)}=-\frac{d(\ln S(t))}{dt}$$

4.1. Introducción

En el capitulo 2 se muestra que las funciones de sobrevida (confiabilidad) y de riesgo pueden estimarse sin necesidad de conocer el modelo probabilístico que genera los datos de tiempo hasta el evento. Esto se hace mediante el estimador de Kaplan-Meier, el de Nelson-Aalen o el suavizador de núcleo, entre otros. A estos procedimientos se les denomina no paramétricos o libres de distribución.

Las funciones de probabilidad, de sobrevida, de confiabilidad y la de riesgo quedan definidas y asociadas con las expresiones

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} = -\frac{d(\ln S(t))}{dt},$$

de donde se concluye que la función de distribución, $f(t)$, determina tanto a $S(t)$ como a $h(t)$. De manera que, si se conoce la forma funcional o la familia distribucional del modelo probabilístico que genera los datos de tiempo para el evento, el problema estadístico consiste en encontrar, a partir de un conjunto de datos observado, el valor de los parámetros que lo define, y así, mediante la expresión anterior, conocer las funciones de sobrevida y de riesgo.

En este capítulo se asume que el modelo probabilístico pertenece a una familia de distribuciones, la cual queda definida por un conjunto de parámetros que, en el contexto frecuentista, son cantidades fijas desconocidas. Así, un modelo probabilístico exponencial queda determinado por un parámetro. Los modelos Weibull, log-normal, log-logístico, gama y Birnbaum-Saunders quedan definidos por dos parámetros, mientras que un modelo gama generalizado queda caracterizado por tres parámetros. Para cada estudio asociado con situaciones, fenómenos o experimentos que involucren la variable *tiempo hasta el evento*, los parámetros de los modelos probabilísticos asumidos deben ser estimados desde los datos muestrales, para así responder a las preguntas que el estudio se proponga responder.

La estrategia adoptada en este capítulo es la construcción de la verosimilitud conjunta de los datos, con censura (Tipo I o Tipo II) y truncamiento como funciones de los parámetros en los modelos de sobrevida. El objetivo es estimar los parámetros mediante la maximización de la función de verosimilitud.

4.2. Construcción de la función de verosimilitud

Suponga una función de probabilidad (discreta o continua) $f(t, \theta)$. En el problema de estimación se conoce un valor particular o realización de la muestra, $\mathbb{T} = (T_1, T_2, \ldots, T_n)$, pero θ es desconocido. Cuando se varía θ manteniendo fija a $\mathbb{T}$, se obtiene la *función de verosimilitud*, $L(\theta|\mathbb{T})$. Esta escritura enfatiza que $L(\cdot)$ es función del parámetro θ para una muestra particular, por lo que simplemente se escribe $L(\theta)$. Es decir, puede escribirse de estas maneras indistintamente:

$$L(\theta|\mathbb{T}) = L(\theta) = f(\mathbb{T}_0|\theta), \ \mathbb{T}_0 \text{ fija}, \theta \text{ variable}.$$

El parámetro θ puede ser un escalar, un vector o una matriz. Para el caso de una distribución exponencial, θ es un escalar. En la distribución lognormal los parámetros son μ y σ), por lo que se dice que es una distribución biparamétrica. Mientras que la distribución gama generalizada tiene tres parámetros, α, θ y γ.

Suponga que $T_1, T_2; \ldots, T_n$ son tiempos hasta un evento independientes e idénticamente distribuidos (iid), con función de probabilidad $f(t; \theta)$, $\theta \in \Omega$; siendo Ω el conjunto donde el parámetro θ toma valores, denominado el *espacio de parámetros*. Notamos de esta manera la función de probabilidad, sea esta para una variable aleatoria discreta o para una variable aleatoria continua; los resultados tienen validez para ambos tipos de variables.

La base del proceso inferencial es la función de verosimilitud dada por

$$L(\theta; t) = f(t_1; \theta) f(t_2; \theta) \cdots f(t_n; \theta) = \prod_{i=1}^{n} f(t_i; \theta), \quad \theta \in \Omega, \qquad (4.2.1)$$

donde $t = (t_1, \ldots, t_n)'$ es una realización de la muestra aleatoria $\mathbb{T}$ (valores observados). Se ha invertido el orden de escritura de t y θ en L para enfatizar que L es función de θ para una muestra dada, t. Como la función logaritmo es una función monótona creciente y continua, entonces se conservan los valores de θ que maximizan L, es decir, los puntos críticos son los mismos para L que para el logaritmo de L; además que se facilitan los cálculos con $\ln L(\cdot)$. Así,

$$l(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(t_i; \theta), \quad \theta \in \Omega. \qquad (4.2.2)$$

Para determinar el estimador máximo verosímil (en adelante **EMV**) de θ, que se nota por $\hat{\theta}_{MV}$, se encuentran los puntos donde la derivada de $l(\theta)$ se

anula, es decir, se debe resolver el sistema de ecuaciones

$$\frac{\partial l(\theta)}{\partial \theta} = 0. \tag{4.2.3}$$

Nuevamente, puede ser un sistema de ecuaciones dependientes de la naturaleza escalar, vectorial e incluso matricial de θ.

No hay garantía de que el sistema de ecuaciones (4.2.3) tenga solución, o así exista una solución, tampoco hay garantía de que esta sea única. Además, otro problema es el procedimiento para resolver (4.2.3), pues no siempre resulta sencillo o inmediato resolver el sistema de ecuaciones. Ante esta última dificultad, se puede optar por el uso de métodos numéricos.

La derivada del logaritmo de la verosimilitud se denomina la función de puntuación (en inglés, *score function*), y se nota por $\mathcal{U}(\theta) = \partial l(\theta)/\partial\theta$.

Para distribuciones cuyo rango de valores posibles es conocido y no depende de ningún parámetro, el procedimiento de máxima verosimilitud proporciona estimadores que son:

1) Asintóticamente insesgados (centrados).
2) Con distribución asintóticamente normal.
3) Asintóticamente de varianza mínima (eficientes).
4) Si existe un estadístico suficiente para el parámetro, el estimador máximo verosímil es suficiente.
5) Invariante, en el sentido de que, si $\hat{\theta}_{MV}$ es el estimador máximo verosímil del parámetro θ, y g es una función continua biunívoca (uno a uno), $g(\hat{\theta}_{MV})$ es el estimador máximo-verosímil de $g(\theta)$.

Se define la *información de Fisher esperada*, $\mathcal{I}(\theta)$, como

$$\mathcal{I}(\theta) = E\left(-\frac{\partial^2 l(\theta)}{\partial\theta_j \partial\theta_k}\right) = E(-H(\theta)), \tag{4.2.4}$$

a la matriz de segundas derivadas parciales de $l(\theta)$ respecto de θ, $H(\theta)$, se le denomina matriz Hessiana.

Se define la información observada como

$$\mathcal{IO}(\hat{\theta}) = \left(-\frac{\partial^2 l(\hat{\theta})}{\partial\hat{\theta}_j \partial\hat{\theta}_k}\right). \tag{4.2.5}$$

Se demuestra que según Hogg, McKean y Craig [32, cap. 6], bajo ciertas condiciones de regularidad, la función de puntuación, $\mathcal{U}(\theta)$, converge en

distribución (distribución asintótica) a una distribución normal con media 0 y varianza $\mathcal{I}(\theta)$, es decir,

$$\mathcal{U}(\theta) \xrightarrow{D} N(0, \mathcal{I}(\theta)). \tag{4.2.6}$$

Nuevamente, los estimadores máximo verosímiles, $\hat{\theta}_{MV}$, se obtienen del sistema (4.2.3). No siempre es sencillo, en términos algebraicos y de cálculo, resolver este sistema de ecuaciones, de manera que se requiere de métodos numéricos con los cuales se puede encontrar una solución exacta o aproximada.

El método de *Newton-Raphson* es una herramienta del análisis numérico útil para este propósito. A continuación se resume su fundamento teórico y se presenta el algoritmo de cálculo del método. Este algoritmo se describe en (4.2.8).

Se asume que la función $g(\theta) = \frac{\partial l(\theta)}{\partial \theta}$ es continuamente diferenciable en una vecindad de la solución buscada. Por aproximación de Taylor alrededor del punto θ_0, sin pérdida de generalidad de manera univariada, se tiene que

$$g(\theta) = g(\theta_0) + (\theta - \theta_0)g'(\theta_0) = 0,$$

de donde

$$\theta = \theta_0 - \frac{g(\theta_0)}{g'(\theta_0)}.$$

A partir de la expresión anterior se puede inducir que

$$\theta^{(m+1)} = \theta^{(m)} - \frac{g(\theta^{(m)})}{g'(\theta^{(m)})} = \theta^{(m)} - \left[g'(\theta^{(m)})\right]^{-1} g(\theta^{(m)}),$$

con $\theta^{(m+1)}$ el valor de θ en el paso $m+1$ y $\theta^{(m)}$ el valor de θ en el paso m, $g(\theta^{(m)})$ la función $g(\theta)$ evaluada en $\theta^{(m)}$ y $g'(\theta^{(m)})$ la derivada de la función $g(\theta)$ evaluada en $\theta^{(m)}$. En Díaz y Morales [21, pp. 599-600] se esquematiza el principio del algoritmo de Newton-Raphson. Se desea encontrar el valor de θ en el cual la función g intersecta ("cruza") el eje θ.

En versión multivariada, es decir, considerando θ como un vector de parámetros, $g'(\hat{\theta}) = \left(\frac{\partial^2 l(\hat{\theta})}{\partial \hat{\theta}_j \partial \hat{\theta}_k}\right)$ equivale a $\mathcal{IO}(\widehat{\theta})$. La expresión recursiva de Newton-Raphson multivariada es

$$\theta^{(m+1)} = \theta^{(m)} + (\mathcal{IO}(\hat{\theta}_{MV})^{-1})^{(m)}(\mathcal{U}(\hat{\theta}_{\mathcal{MV}}))^{(m)} \tag{4.2.7}$$

Cuando se dispone de la matriz de información de Fisher Esperada (10.3.15), el algoritmo de Newton-Raphson en (4.2.7) toma la forma

$$\theta^{(m+1)} = \theta^{(m)} + (\mathcal{I}(\theta)^{-1})^{(m)}(\mathcal{U}(\theta))^{(m)}. \tag{4.2.8}$$

El proceso se termina cuando el valor de θ^{m+1} converge (se "estabiliza") para algún m entero positivo; es decir, cuando $|\theta^{(m+1)} - \theta^{(m)}| < \epsilon$, para $\epsilon > 0$. Se toma como solución del sistema de ecuaciones este valor de θ en el paso $m + 1$.

Bajo ciertas condiciones de regularidad [32, cap. 6] y por las propiedades de los EMV enunciadas arriba, la distribución asintótica de $\hat{\theta}_{MV}$ es

$$\hat{\theta}_{MV} \xrightarrow{D} N(\theta, \mathcal{I}(\theta)^{-1}). \tag{4.2.9}$$

Además de la normalidad asintótica de $\hat{\theta}_{MV}$, se concluye que es asintóticamente centrado (insesgado) y que su varianza asintótica es el inverso de la información esperada de Fisher. Se puede estimar la varianza asintótica del EMV $\hat{\theta}$ de θ como

$$\widehat{Var}(\hat{\theta}_{MV}) \approx \mathcal{I}O(\hat{\theta}_{MV})^{-1}, \tag{4.2.10}$$

es decir, la inversa de la información observada.

Con los elementos anteriores se puede hacer inferencia sobre θ en términos de la construcción de intervalos o regiones de confianza para este parámetro, o en términos de la verificación de hipótesis sobre θ.

Intervalo de confianza

En caso de que θ sea un escalar, por los resultados contenidos en las expresiones (4.2.9) y (4.2.10), un intervalo de $(1-\alpha)100\,\%$ de confianza para θ es dado por

$$\hat{\theta}_{MV} \mp Z_{1-\alpha/2} ee(\hat{\theta}_{MV}), \tag{4.2.11}$$

donde $ee(\hat{\theta})$ es el error estándar del estimador, este es igual a

$$ee(\hat{\theta}) = \sqrt{\widehat{Var}(\hat{\theta}_{MV})} \approx \sqrt{\mathcal{I}O(\hat{\theta}_{MV})^{-1}}. \tag{4.2.12}$$

Para el caso de que θ sea un vector de tamaño $p \times 1$, la estadística dada por

$$\chi^2 = (\hat{\theta}_{MV} - \theta)' \mathcal{I}(\theta)(\hat{\theta}_{MV} - \theta) \tag{4.2.13}$$

tiene una distribución asintótica ji-cuadrado con p grados de libertad.

La expresión (4.2.13) es equivalente con

$$\chi^2 = (\theta - \hat{\theta}_{MV})' \mathcal{I}(\theta)(\theta - \hat{\theta}_{MV}),$$

y corresponde a la distancia de Mahalanobis entre el vector de parámetros θ y su EMV, $\hat{\theta}_{MV}$. Desde un punto de vista geométrico, esta expresión representa una elipsoide con centro en $\hat{\theta}_{MV}$ y distancia $\chi^2_{(1-\alpha,p)}$. Por la distribución de esta forma cuadrática se puede hacer que

$$P\left[(\hat{\theta}_{MV} - \theta)' \mathcal{I}(\theta)(\hat{\theta}_{MV} - \theta) \leq \chi^2_{(1-\alpha,p)}\right] = 1 - \alpha,$$

desde la cual una *región de confianza* para θ está conformada por todos los vectores θ que satisfacen

$$(\theta - \hat{\theta}_{MV})' \mathcal{I}(\theta)(\theta - \hat{\theta}_{MV}) \leq \chi^2_{(1-\alpha,p)}. \tag{4.2.14}$$

Verificación de hipótesis

Para verificar la hipótesis $H_0 : \theta = \theta_0$ (θ un escalar), se recurre a la distribución asintótica de la función de puntuación $U(\hat{\theta})$ o la del EMV $\hat{\theta}$. Se dispone de las siguientes tres estadísticas para la verificación o la contrastación de H_0:

1. Estadística de Wald: $(\hat{\theta} - \theta_0)^2 \mathcal{I}(\theta) \approx \left(\frac{\hat{\theta}-\theta_0}{ee(\hat{\theta})}\right)^2 \xrightarrow{D} \chi^2_{(1)}$
2. Razón de verosimilitud: $-2\ln\left(\frac{L(\theta_0)^2}{L(\hat{\theta})}\right) \xrightarrow{D} \chi^2_{(1)}$
3. Estadística de puntuación: $\frac{U(\theta_0)^2}{\mathcal{I}(\theta_0)} \xrightarrow{D} \chi^2_{(1)}$

Cuando la información de Fisher esperada, $\mathcal{I}(\hat{\theta})$, no se puede obtener, entonces se opta por la información observada, $\mathcal{I}O(\hat{\theta})$, la cual no altera la distribución asintótica de las tres estadísticas anteriores para la verificación de H_0.

La verificación de hipótesis sobre el parámetro θ, cuando este es un vector de tamaño $p \times 1$, se hace de manera similar a como se procede para la construcción de regiones de confianza de θ. En este sentido, la estadística de verificación tiene la forma contenida en (4.2.13), con la cual la estadística para verificar $H_0 : \theta = \theta_0$ es

$$\chi^2_0 = (\hat{\theta}_{MV} - \theta_0)' \mathcal{I}(\theta)(\hat{\theta}_{MV} - \theta_0) \xrightarrow{D} \chi^2_{(p)}. \tag{4.2.15}$$

A un nivel de significación $(1-\alpha)100\,\%$, los valores de χ^2_0 superiores $\chi^2_{(1-\alpha,p)}$ sugieren un rechazo de la hipótesis nula $H_0 : \theta = \theta_0$. Note que cuando $p = 1$ (θ escalar), la estadística de Wald anterior es un caso especial de la estadística (4.2.15).

4.2.1. Verosimilitud para datos con censura Tipo I

Una de las características casi exclusivas de los datos de tiempo para un evento es la presencia de censura y truncamiento, los cuales deben considerarse e incluirse en la función de verosimilitud. Un supuesto importante es que los tiempos al evento y los tiempos de censura son independientes. En la construcción de la función de verosimilitud para datos censurados o truncados

(estos dos términos no son lo mismo), se requiere establecer la información que cada uno de los datos, sean exactos, censurados o truncados, suministra a la verosimilitud.

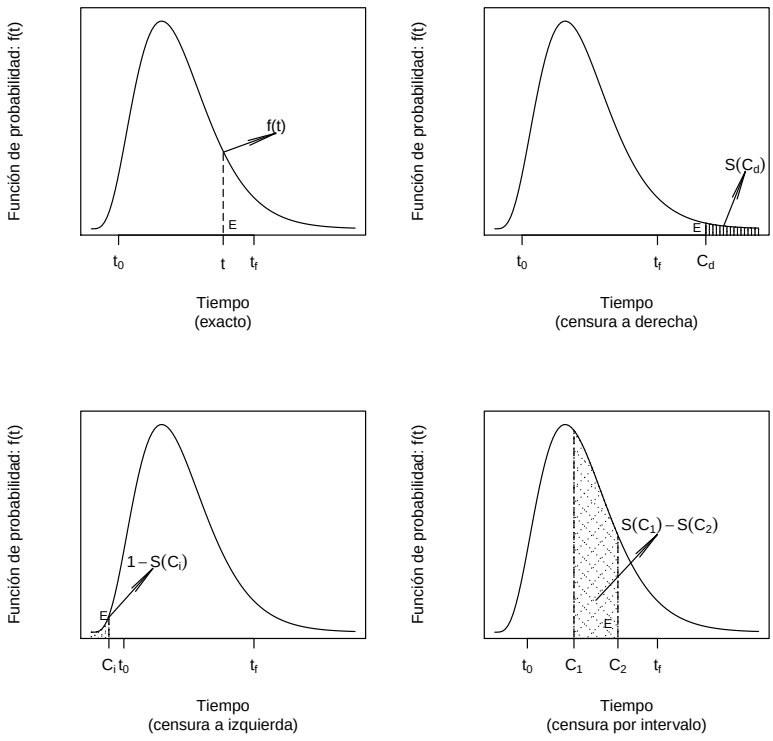

Figura 4.1. Información de probabilidad para la verosimilitud (evento E).

Una observación asociada con un dato de tiempo al evento exacto suministra información sobre la probabilidad de ocurrencia del evento en ese tiempo, la cual es aproximadamente igual a la función de probabilidad (discreta o continua) de la variable aleatoria **T** en este tiempo. Para una observación con censura a la derecha, la información disponible es que el tiempo al evento es mayor que el tiempo al momento de la censura, por lo tanto, la información probabilística es la función de sobrevida evaluada sobre el momento de la censura a derecha (C_d).

Para datos con censura a derecha Tipo I, suponga que los tiempos al evento $T_1, T_2, \ldots, T_n$ son censurados a derecha por las constantes $c_1, \ldots, c_n$. Estos datos pueden escribirse como pares ordenados (Z_i, δ_i), para

$i = 1, 2, \ldots, n$, donde $Z_i = \text{mín}\{T_i, c_i\}$ y

$$\delta_i = \begin{cases} 1, & \text{si } T_i \leq c_i \text{ (no censurado)} \\ 0, & \text{si } T_i > c_i \text{ (censurado)}. \end{cases}$$

Para la construcción de la verosimilitud se ordenan los datos observados del menor al mayor, del mismo modo se ordenan los respectivos indicadores de censura; es decir, se dispone $T_{(1)} \leq T_{(2)} \leq, \ldots, \leq T_{(n)}$ con $\delta_{(i)} = \delta_j$ cuando $Z_{(i)} = Z_j$. El objetivo es determinar la verosimilitud de los datos (Z_i, δ_i), para $i = 1, 2, \ldots, n$, de manera que el parámetro θ pueda ser estimado mediante el método de máxima verosimilitud y, a partir de esta estimación, se procede a desarrollar la respectiva inferencia sobre θ, es decir, la verificación de hipótesis y la construcción de intervalos de confianza.

En el libro de Smith [64] se muestra que bajo censura a derecha Tipo I con tiempos de censura fijos, la función de verosimilitud, $L(\theta)$, de los datos observados (Z_i, δ_i), para $i = 1, 2, \ldots, n$, es

$$L(\theta) = \prod_{i=1}^{n} f(z_i)^{\delta_i} S(z_i)^{1-\delta_i}. \tag{4.2.16}$$

De acuerdo con la expresión (1.5.2), $f(t) = h(t)S(t)$, y por la igualdad (1.5.4), $S(T) = \exp(-H(t))$, la verosimilitud se puede escribir como

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} (h(z_i)S(z_i))^{\delta_i} S(z_i)^{1-\delta_i} \\ &= \prod_{i=1}^{n} h(z_i)^{\delta_i} S(z_i) \\ &= \prod_{i=1}^{n} h(z_i)^{\delta_i} \exp[-H(z_i)] \end{aligned} \tag{4.2.17}$$

De manera análoga, para una observación con censura a izquierda todo lo que se conoce es que el evento ya ha ocurrido, por tanto, la contribución de esta observación a la verosimilitud es la función de distribución acumulada evaluada en el tiempo de la censura a izquierda (C_i). Para una observación con censura por intervalo se conoce que el evento ha ocurrido dentro del intervalo, en consecuencia, la información conocida es la probabilidad de que el tiempo al evento esté dentro de este intervalo. Para datos truncados, estas probabilidades se reemplazan por las probabilidades condicionales pertinentes.

La figura 4.1 y la tabla 4.1 ilustran los componentes de probabilidad de la función de verosimilitud de acuerdo con la censura Tipo I.

Tabla 4.1. Componentes de probabilidad bajo censura Tipo I

Tipo de tiempo	Componente de probabilidad
Tiempo exacto	$f(t)$
Censura a derecha	$S(C_d)$
Censura a izquierda	$1 - S(C_i) = F(C_i)$
Censura por intervalo	$S(C_1) - S(C_2)$

Así, la función de verosimilitud para estos casos de censura o truncamiento debe considerar e incorporar los siguientes componentes probabilísticos.

Para el caso de datos con censura Tipo I, la función de verosimilitud se construye poniendo juntos estos componentes:

$$L = \prod_{[j \in \mathscr{E}]} f(t_j) \prod_{j \in \mathscr{D}} (S(C_{d_j})) \prod_{j \in \mathscr{I}} (1 - S(i_j)) \prod_{j \in \mathscr{I}_n} (S(C_{1_j}) - S(C_{2_j}),$$

donde $\mathscr{E}$ es el conjunto de unidades a las cuales les ha ocurrido el evento E, el conjunto $\mathscr{D}$ representa el conjunto de unidades con censura a derecha, $\mathscr{I}$ es el conjunto de unidades con censura a izquierda e $\mathscr{I}_n$ es el conjunto conformado por las unidades con censura por intervalo. Naturalmente, estos conjuntos son disjuntos e incluso vacíos si algún tipo de censura no se presenta.

4.2.2. Verosimilitud para datos truncados

Se reitera que, en el truncamiento, las partes de la distribución de la variable aleatoria *tiempo hasta la ocurrencia del evento*, **T**, que se consideran son solo aquellas cuyos valores se encuentren por encima (o por debajo) del punto de truncamiento.

Así, en un estudio se puede estar interesado en los tiempos hasta el evento cuyo valor sea superior a un valor T_i determinado (para todas las unidades o para cada unidad); en consecuencia, las unidades cuyos tiempos al evento estén por debajo de T_i quedan excluidas del estudio o son *truncadas a izquierda*. Una caso similar se tiene en situaciones en las cuales solo se consideran

los tiempos hasta el evento, siempre que estos sean menores que un tiempo T_d; las unidades que no cumplan esta condición son excluidas o quedan *truncadas a derecha*. Para el caso de observaciones *truncadas por intervalo* se deben cumplir conjuntamente las condiciones de truncamiento a izquierda y a derecha; es decir, se excluyen o truncan las unidades que estén por fuera de la ventana $[T_i, T_d]$. Para datos truncados a izquierda, con T_i siendo el tiempo de truncamiento y asumiendo que hay independencia entre los tiempos al evento y el mecanismo que genera el truncamiento constante, la densidad condicional se puede escribir como

$$f(t|t \geq T_i) = \frac{f(t)}{S(T_{ij})}.$$

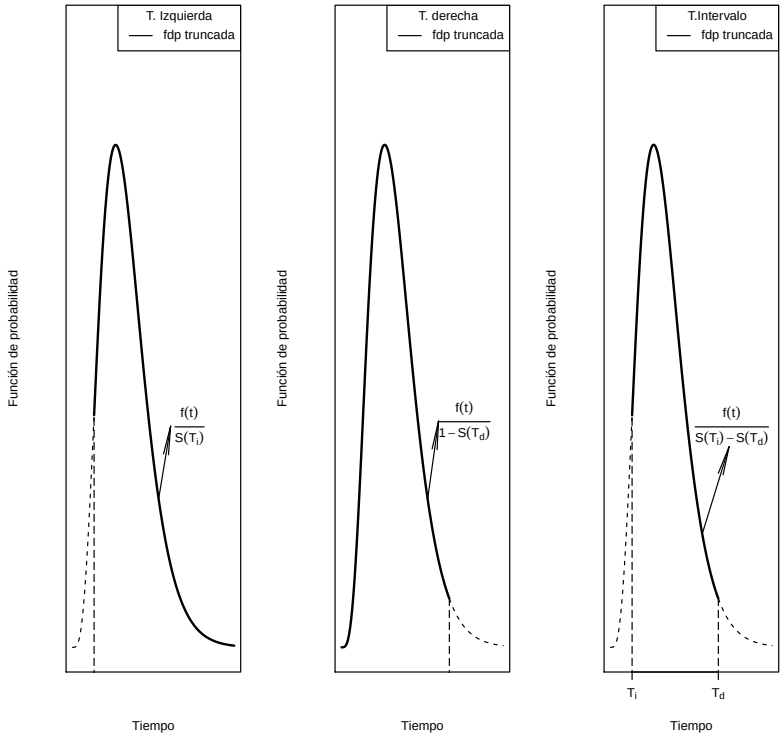

Figura 4.2. Funciones de probabilidad truncadas.

En consecuencia, $f(t)$ se reemplaza por $f(t)/S(T_i))$. Por lo tanto, la contribución a la verosimilitud es

$$L = \prod_{j=1}^{n} \frac{f(t_j)}{S(T_{ij})}.$$

Una argumentación similar se tiene para datos con truncamiento a derecha, esto es

$$f(t|t \leq T_d) = \frac{f(t)}{F(T_d)} = \frac{f(t)}{1 - S(T_d)}.$$

Así, la verosimilitud es

$$L = \prod_{j=1}^{n} \frac{f(t_j)}{1 - S(T_{dj})}.$$

Tabla 4.2. Función de probabilidad bajo truncamiento

Tipo de tiempo	Componente de probabilidad
Truncamiento a izquierda	$f(t)/S(T_i)$
Truncamiento a derecha	$f(t)/(1 - S(T_d))$
Truncamiento por intervalo	$f(t)/(S(T_i) - S(T_d))$

Finalmente, para datos con truncamiento de intervalo, la función de probabilidad (discreta o continua), $f(t)$, igualmente queda truncada, de manera que la variable aleatoria $\mathbf{T}$ tiene distribución $f(t)/[S(T_i)) - S(T_d)]$, para $\mathbf{T}$ definida entre T_i y T_d. En la figura 4.2 y en la tabla 4.2 se muestran los elementos de probabilidad que se deben considerar e incluir en la respectiva función de verosimilitud.

4.2.3. Verosimilitud para datos con censura Tipo II

En el esquema de censura Tipo II solo se consideran los r tiempos menores hasta la ocurrencia del evento, $T_{(1)} \leq T_{(2)} \leq, \ldots, \leq T_{(r)}$, a partir de una muestra de $n \geq r$ tiempos $T_1, T_2, \ldots, T_n$. En estas condiciones la función de verosimilitud para datos con censura Tipo II es proporcional a la distribución conjunta de las primeras r estadísticas, $T_{(1)}, T_{(2)}, \ldots, T_{(r)}$. Por

teoría estadística, la función de densidad conjunta de estas estadísticas de orden [32] es

$$f(t_{(1)}, t_{(2)}, \ldots, t_{(r)}) = \frac{n!}{(n-r)!} f(t_{(1)}) f(t_{(2)}) \cdots f(t_{(r)}) S(t_{(r)})^{n-r}. \tag{4.2.18}$$

La función de verosimilitud es

$$L(\theta) \propto f(t_{(1)}) f(t_{(2)}) \cdots f(t_{(r)}) S(t_{(r)})^{n-r},$$

donde $\propto$ significa proporcional. Se observa la similitud de la función de verosimilitud para esta estructura de datos con la verosimilitud de datos con censura Tipo I, pues allí las observaciones completas se evalúan en la función de probabilidad ($f(t)$) y las observaciones censuradas son evaluadas en la función de sobrevida $S(t)$.

4.2.4. Verosimilitud para datos con censura aleatoria

Como se describe en el capítulo 1, en este mecanismo de censura se asume que el tiempo de censura C_i para la i-ésima unidad es una variable aleatoria. Se asume que los C_i son independientes de los tiempos al evento $T_1, T_2, \ldots, T_n$. En esta notación, C_i es el censor aleatorio de tiempo asociado con T_i y es independiente de este, para $i = 1, 2, \ldots, n$.

En consecuencia, las $C_1, C_2, \ldots, C_n$ son variables aleatorias independientes e idénticamente distribuidas (iid), las cuales tienen una función de probabilidad f_C y una función de sobrevida S_C. Mientras que los tiempos al evento, $T_1, T_2, \ldots, T_n$, son iid con función de probabilidad $f_T(\theta)$ y función de sobrevida S_T.

Los datos con censura aleatoria a derecha se conforman con la siguiente estructura: por los pares (Z_i, δ_i) para $i = 1, \ldots, n$, donde $Z_i = \text{mín}\{T_i, C_i\}$ y

$$\delta_i = \begin{cases} 1, & \text{si } T_i \leq C_i \text{ (no censurado)} \\ 0, & \text{si } T_i > C_i \text{ (censurado).} \end{cases} \tag{4.2.19}$$

El propósito es determinar la función de verosimilitud de los datos observados y censurados de la forma (Z_i, δ_i), $i = 1, 2, \ldots, n$, tal que el parámetro θ pueda ser estimado maximizando la verosimilitud.

Por los supuestos anteriores de independencia, identificación distribucional y de sobrevida, la verosimilitud se factoriza en dos partes, una relacionada con los tiempos de censura aleatoria y otra relacionada con los

tiempos al evento [37]. Esta es

$$L(\theta) = c\left(\prod_{i=1}^{n} S_C(z_i)^{\delta_i} f_C(z_i)^{1-\delta_i}\right)\left(\prod_{i=1}^{n} f_T(z_i)^{\delta_i} S_T(z_i)^{1-\delta_i}\right). \qquad (4.2.20)$$

Note que en (4.2.20) el primer factor contiene los componentes de probabilidad para los tiempos de censura aleatoria. Si no hay parámetros de interés en f_C y S_C (es decir, no dependen de θ), entonces este factor se considera como una constante multiplicativa. La expresión de verosimilitud (4.2.20) coincide con la verosimilitud (4.2.16).

Se debe resaltar que el supuesto de independencia entre los tiempos al evento y la censura aleatoria, T y C, no siempre se tiene. En el ejemplo de las mujeres con cáncer en ovarios puede resultar que el progreso de los resultados del estudio provoque una terminación anticipada del mismo en el caso de algunas mujeres; es decir, un bienestar en la salud debido al estudio induce un retiro del mismo, y esto altera el supuesto de independencia. Otro ejemplo se tiene con los trasplantes de órganos, pues en estudios que incluyen pacientes jóvenes, estos tienden a abandonar el mismo; es decir que la edad tiene un efecto significativo en la última sobrevida observada o registrada, y esto hace que el tiempo al evento y la censura no sean independientes. Como conclusión, se debe tener claro el protocolo, las unidades de observación y el desarrollo del estudio para establecer si el supuesto de independencia es admisible o no. Esta es una advertencia transversal a la mayoría de las aplicaciones de los métodos estadísticos.

Estimación del parámetro en el modelo exponencial

Datos sin censura

Suponga que n unidades son seguidas hasta que presentan un evento específico. Se asume que la variable aleatoria $\mathbf{T}$ tiene una distribución exponencial con parámetro θ, $\mathbf{T} \sim \exp(\theta)$. Sea $t_1, t_2, \ldots, t_n$ los tiempos observados exactos al evento de las n unidades. La función de verosimilitud, usando (4.2.16), con $\delta_i = 1$, para $i = 1, 2, \ldots, n$, es

$$L(\theta) = \prod_{i=1}^{n} \theta e^{-\theta t_i},$$

la función log-verosimilitud es

$$l(\theta) = n \ln \theta - \theta \sum_{i=1}^{n} t_i.$$

El estimador de máxima verosimilitud de θ se obtiene igualando a 0 esta última expresión y resolviendo para θ, este es

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} t_i}. \tag{4.2.21}$$

Dado que la media de la distribución exponencial es $\mu = 1/\theta$ y que los estimadores de máxima verosimilitud (EMV) son invariantes bajo transformaciones uno a uno, el EMV de μ es

$$\hat{\mu} = \frac{1}{\hat{\theta}} = \frac{\sum_{i=1}^{n} t_i}{n} = \bar{t}.$$

4.2.4.1. Datos con censura Tipo I

Suponga que los tiempos $T_1, T_2, \ldots, T_n$ están censurados a derecha por las constantes $c_1, c_2, \ldots, c_n$. La función de verosimilitud para el modelo exp(θ), usando (4.2.16), con $\delta_i = 1$ si la observación es no censurada (11.2.4) y $Z_i = \min\{T_i, C_i\}$, $i = 1, 2, \ldots, n$, es

$$\begin{aligned} L(\theta) &= c \prod_{i=1}^{n} f(z_i)^{\delta_i} S(z_i)^{1-\delta_i} \\ &= c \prod_{i=1}^{n} \left(\theta e^{-\theta z_i}\right)^{\delta_i} \left(e^{-\theta z_i}\right)^{1-\delta_i} \\ &= c\theta^{\sum_{i=1}^{n} \delta_i} e^{-\theta \sum_{i=1}^{n} z_i} \\ &= c\theta^{R} e^{-(\theta \mathcal{T})}, \end{aligned}$$

donde $\mathcal{R} = \sum_{i=1}^{n} \delta_i$ es una variable aleatoria que cuenta el número de tiempos exactos o no censurados y $\mathcal{T} = \sum_{i=1}^{n} Z_i$ representa el total de tiempos observados.

Resolviendo $\frac{d \ln L(\theta)}{d\theta} = \frac{dl(\theta)}{d\theta} = 0$, se encuentra que el EMV de θ es

$$\hat{\theta} = \frac{\mathcal{R}}{\mathcal{T}} = \frac{\mathcal{R}}{\sum_{i=1}^{n} Z_i}.$$

La distribución asintótica de $\hat{\theta}$ es

$$\hat{\theta} \xrightarrow{D} N(\theta, \mathcal{I}(\theta)^{-1}),$$

con error estándar (ee) estimado desde la varianza asintótica

$$ee(\hat{\theta}) = \sqrt{\widehat{Var}(\hat{\theta}_{MV})} \approx \sqrt{\mathcal{I}O^{-1}(\hat{\theta}_{MV})}.$$

Para este caso se nota que la varianza debe ser estimada a través de la información observada, en lugar de la información esperada, de Fisher, desde la expresión

$$\mathcal{I}O(\hat{\theta}) = \frac{\mathcal{R}}{\hat{\theta}^2},$$

porque no todos los tiempos de censura $c_1, c_2, \ldots, c_n$ son conocidos en el esquema de censura Tipo I.

Verosimilitud bajo censura Tipo II

Asuma que los tiempos $T_1, T_2, \ldots, T_n$ son variables aleatorias independientes e idénticamente distribuidas conforme a $\exp(\theta)$, las cuales están sujetas a censura Tipo II. En esta estructura de censura únicamente se observan los tiempos $T_{(1)}, T_{(2)}, \ldots, T_{(r)}$ para un número prefijado de valores $r \leq n$. En la verosimilitud expresada en (4.2.18), la función de probabilidad es $f(t) = \theta e^{-\theta t}$ y la función de sobrevida es $S(t) = e^{-\theta t}$, así

$$\begin{aligned} L(\theta) &= \frac{n!}{(n-r)!} f(t_{(1)} f(t_{(2)}) \cdots f(t_{(r)}) S(t_{(r)})^{n-r} \\ &= \frac{n!}{(n-r)!} \left(\prod_{i=1}^{r} \theta e^{-\theta t_{(i)}} \right) \left(e^{-\theta t_{(i)}} \right)^{n-r} \\ &= \frac{n!}{(n-r)!} \theta^r \exp\left[-\theta \left(\sum_{i=1}^{r} t_{(i)} + (n-r) t_{(r)} \right) \right]. \end{aligned}$$

El EMV de θ se obtiene como solución de $\frac{dl(\theta)}{d\theta} = 0$, este es

$$\hat{\theta} = \frac{\mathcal{R}}{\mathcal{T}}, \tag{4.2.22}$$

donde

$$\mathcal{T} = \sum_{i=1}^{r} t_{(i)} + (n-r) t_{(r)},$$

a $\mathcal{T}$ se le denomina el *tiempo al evento total*. Note que los tiempos censurados y no censurados son sumados, lo que da como resultado $\mathcal{T}$.

La media del tiempo al evento es $\mu = 1/\theta$, entonces, puede ser estimada por

$$\mu = \frac{1}{\hat{\theta}} = \frac{\sum_{i=1}^{r} t_{(i)} + (n-r) t_{(r)}}{r}.$$

Así se demuestra que $2r\theta/\hat{\theta}$ tiene distribución ji-cuadrado con $2r$ grados de libertad [64]. Es decir,

$$\frac{2r\theta}{\hat{\theta}} = 2\theta\left(\sum_{i=1}^{r} t_{(i)} + (n-r)t_{(r)}\right) \sim \chi^2_{(2r)}.$$

Por lo tanto, un intervalo del $(1-\alpha)100\,\%$ de confianza para θ, se construye así:

$$P\left(\chi^2_{(2r,\alpha/2)} < \frac{2r\theta}{\hat{\theta}} < \chi^2_{(2r,1-\alpha/2)}\right) = 1-\alpha$$
$$P\left(\frac{\hat{\theta}\,\chi^2_{(2r,\alpha/2)}}{2r} < \theta < \frac{\hat{\theta}\,\chi^2_{(2r,1-\alpha/2)}}{2r}\right) = 1-\alpha. \qquad (4.2.23)$$

Un intervalo del $(1-\alpha)100\,\%$ de confianza para el parámetro θ es

$$\left[\frac{\hat{\theta}\,\chi^2_{(2r,\alpha/2)}}{2r}\,,\ \frac{\hat{\theta}\,\chi^2_{(2r,1-\alpha/2)}}{2r}\right]. \qquad (4.2.24)$$

Suponga que en un laboratorio experimental 10 ratones son expuestos a carcinógenos [44]. El experimentador decide terminar el estudio después de que la mitad de los ratones mueren y sacrifica la otra mitad en ese momento. Los tiempos de sobrevida de los ratones son 4, 5, 8, 9 y 10 semanas. Los datos de sobrevida de los 10 ratones son 4, 5, 8, 9, 10, 10^+, 10^+, 10^+ y 10^+. Asumiendo que los tiempos a la muerte de estos ratones siguen una distribución exponencial, la tasa de sobrevida θ y la media μ son estimadas mediante (4.2.22), respectivamente.

En este caso,

$$\hat{\theta} = \frac{\mathcal{R}}{\mathcal{T}} = \frac{\sum_{i=1}^{5}\delta_i}{\sum_{i=1}^{5} t_{(i)} + (n-r)t_{(r)}} = \frac{5}{36+50} = 0.058,$$

la media estimada es $\hat{\mu} = 1/0.058 = 17.241$ semanas. Un intervalo del 95 % de confianza para θ estimado mediante (4.2.24) es

$$\frac{(0.058)(3.247)}{(2)(5)} < \theta < \frac{(0.058)(20.483)}{(2)(5)},$$

es decir, $[0.019\ ,\ 0.119]$ es un intervalo del 95 % para el parámetro θ, la tasa de sobrevida.

Estimación de parámetros en el modelo Weibull

La distribución de Weibull tiene función de densidad y de sobrevida (confiabilidad):

$$f(t) = \theta\alpha(\theta t)^{\alpha-1}e^{-(\theta t)^{\alpha}}I_{[0,\infty)}(t)$$
$$S(t) = e^{-(\theta t)^{\alpha}}.$$

Se muestra el procedimiento de estimación e inferencia de los parámetros de un modelo Weibull mediante máxima verosimilitud para los esquemas de censura Tipo I y Tipo II.

4.2.4.2. Datos sin censura (exactos)

Sean $t_1, t_2, \ldots t_n$ los tiempos exactos para la ocurrencia de un evento de n unidades de un estudio. Asumiendo que estos tiempos proceden de una distribución *Weibull*(θ, α), $\delta_i = 1$ para todo $i = 1, 2, \ldots, n$, la función de verosimilitud es

$$L(\theta, \alpha) = f(t_1, t_2, \ldots t_n, \theta, \alpha) = \prod_{i=1}^{n} f(t_i, \theta, \alpha) \tag{4.2.25}$$

$$= \prod_{i=1}^{n} \left(\theta\alpha(\theta t_i)^{\alpha-1}\exp\{-(\theta t_i)^{\alpha}\}\right) \tag{4.2.26}$$

$$= \theta^{n\alpha}\alpha^{n}\exp\left(-\theta\alpha\sum_{i=1}^{n} t_i\right)\prod_{i=1}^{n} t_i^{\alpha-1}. \tag{4.2.27}$$

El logaritmo de esta función de verosimilitud es

$$l(\theta, \alpha) = L(\theta, \alpha) \tag{4.2.28}$$

$$= n\ln\alpha + n\alpha\ln\theta + \sum_{i=1}^{n}\left[(\alpha-1)\ln t_i - \theta^{\alpha}{t_i}^{\alpha}\right]. \tag{4.2.29}$$

Los estimadores de máxima verosimilitud para θ y α se obtienen como la solución del siguiente sistema de ecuaciones:

$$U(\theta) = \frac{\partial l(\theta, \alpha)}{\partial\theta} = n - \theta^{\alpha}\sum_{i=1}^{n} t_i = 0$$
$$U(\alpha) = \frac{\partial l(\theta, \alpha)}{\partial\alpha} = \frac{n}{\alpha} - n\ln\theta + \sum_{i=1}^{n}\ln t_i - \theta^{\alpha}\sum_{i=1}^{n} t_i^{\alpha}(\ln\theta + \ln t_i) = 0. \tag{4.2.30}$$

Si α es conocido, el EMV de $\hat{\theta}$ de θ es

$$\hat{\theta}_\alpha = \left(\frac{n}{\sum_{i=1}^n t_i^\alpha} \right)^{1/\alpha}.$$

Un procedimiento para encontrar una solución aproximada a este sistema de ecuaciones es el método de Newton-Raphson.

4.2.4.3. Datos con censura Tipo I

Como en la distribución exponencial, sean los tiempos $T_1, T_2, \ldots, T_n$, censurados a derecha por las constantes $c_1, c_2, \ldots, c_n$. La función de verosimilitud para el modelo *Weibull*(θ, α), empleando (4.2.16), con $\delta_i = 1$ si la observación es completa ((11.2.4)), $i = 1, 2, \ldots, n$, es

$$\begin{aligned} L(\theta, \alpha) &= \prod_{i=1}^n f(z_i, \theta, \alpha)_i^\delta S(z_i, \theta, \alpha)^{1-\delta_i} \\ &= \prod_{i=1}^n \left[\theta\alpha(\theta z_i)^{\alpha-1} \exp\{-(\theta z_i)^\alpha\} \right]^{\delta_i} \left[\exp\{-(\theta z_i)^\alpha\} \right]^{1-\delta_i} \\ &= \theta^{\sum_{i=1}^n \delta_i} \alpha^{\sum_{i=1}^n \delta_i} \prod_{i=1}^n (\theta z_i)^{(\alpha-1)\delta_i} \exp\{-(\theta z_i)^\alpha\}. \end{aligned}$$

El logaritmo de $L(\theta, \alpha)$ es

$$\begin{aligned} l(\theta, \alpha) &= \ln L(\theta, \alpha) \\ &= \sum_{i=1}^n \delta_i \ln \alpha + \alpha \sum_{i=1}^n \delta_i \ln \theta + (\alpha - 1) \sum_{i=1}^n \delta_i \ln z_i - \theta^\alpha \sum_{i=1}^n z_i^\alpha \\ &= \mathcal{R} \ln \alpha + \alpha \mathcal{R} \ln \theta + (\alpha - 1) \sum_{i=1}^n \delta_i \ln z_i - \theta^\alpha \sum_{i=1}^n z_i^\alpha. \end{aligned} \tag{4.2.31}$$

Las funciones de puntuación, en θ y α, respectivamente, son

$$\begin{aligned} U(\theta) &= \frac{\partial l(\theta, \alpha)}{\partial \theta} = \frac{\alpha \mathcal{R}}{\theta} - \alpha \theta^{\alpha-1} \sum_{i=1}^n z_i^\alpha \\ U(\alpha) &= \frac{\partial l(\theta, \alpha)}{\partial \alpha} = \frac{\mathcal{R}}{\alpha} + \mathcal{R} \ln \theta + \sum_{i=1}^n \delta_i \ln z_i - \theta^\alpha \sum_{i=1}^n z_i^\alpha \ln(\theta z_i). \end{aligned} \tag{4.2.32}$$

Al igualar a 0 tanto $U(\theta)$ como $U(\alpha)$, y al resolver este sistema de ecuaciones respecto a θ y α, se obtienen los respectivos EMV.

Si se asume que α es conocido, el EMV de θ es

$$\hat{\theta}_\alpha = \left(\frac{\mathcal{R}}{\sum_{i=1}^n z_i^\alpha} \right)^{1/\alpha}.$$

Este resultado también podría obtenerse del hecho de que si la variable aleatoria $\mathbf{T} \sim Weibull(\theta, \alpha)$, entonces la variable aleatoria $\mathbf{T}^\alpha$ tiene distribución exponencial con parámetro θ^α.

Sustituyendo el EMV $\hat{\theta}_\alpha$ en la ecuación $U(\alpha) = 0$, se obtiene, después de algunas simplificaciones algebraicas, la ecuación

$$\frac{\mathcal{R}}{\alpha} + \sum_{i=1}^n \delta_i \ln z_i - \frac{\mathcal{R} \sum_{i=1}^n z_i^\alpha \ln z_i}{\sum_{i=1}^n z_i^\alpha} = 0.$$

La solución de esta ecuación (que no contiene a θ) corresponde al EMV $\hat{\alpha}$. El sistema de ecuaciones anterior puede resolverse de manera iterativa a través de métodos numéricos como el de Newton-Raphson.

La matriz de información observada, $\mathcal{I}O(\hat{\theta}, \hat{\alpha})$, se obtiene de las segundas derivadas evaluadas en los EMV $\hat{\theta}$ y $\hat{\alpha}$ con signo opuesto. Estas derivadas son

$$\begin{aligned}
\frac{\partial^2 l(\theta, \alpha)}{\partial \theta^2} &= -\frac{\alpha \mathcal{R}}{\theta^2} - \alpha(\alpha - 1)\theta^{\alpha-2} \sum_{i=1}^n z_i^\alpha \\
\frac{\partial^2 l(\theta, \alpha)}{\partial \theta \partial \alpha} &= \frac{\mathcal{R}}{\theta} - \theta^{\alpha-1}(1 + \alpha \ln \theta) \sum_{i=1}^n z_i^\alpha - \alpha \theta^{\alpha-1} \sum_{i=1}^n z_i^\alpha \ln z_i \\
\frac{\partial^2 l(\theta, \alpha)}{\partial \alpha^2} &= -\frac{\mathcal{R}}{\alpha^2} - \theta^\alpha \sum_{i=1}^n z_i^\alpha [\ln(\theta z_i)]^2.
\end{aligned} \tag{4.2.33}$$

La *matriz de covarianzas de los EMV*, $Var(\hat{\beta}, \hat{\alpha})$, se obtiene como la inversa de la *matriz de información observada*, $\mathcal{I}O(\hat{\theta}, \hat{\alpha})$. Esta es:

$$\begin{aligned}
Var(\hat{\beta}, \hat{\alpha}) &= \mathcal{I}O(\hat{\theta}, \hat{\alpha})^{-1} \\
&= \begin{pmatrix} -\frac{\partial^2 l(\theta,\alpha)}{\partial \theta^2}\Big|_{\theta=\hat{\theta}} & -\frac{\partial^2 l(\theta,\alpha)}{\partial \theta \partial \alpha}\Big|_{(\theta,\alpha)=(\hat{\theta},\hat{\alpha})} \\ -\frac{\partial^2 l(\theta,\alpha)}{\partial \theta \partial \alpha}\Big|_{(\theta,\alpha)=(\hat{\theta},\hat{\alpha})} & -\frac{\partial^2 l(\theta,\alpha)}{\partial \alpha^2}\Big|_{\alpha=\hat{\alpha}} \end{pmatrix}^{-1}.
\end{aligned} \tag{4.2.34}$$

Las entradas de esta matriz están en las expresiones contenidas en (4.2.33), a las cuales se les debe cambiar el signo.

Con la inversa de la matriz de información observada o los elementos de su diagonal (las varianzas observadas) se pueden construir regiones e intervalos de confianza o desarrollar la verificación de hipótesis sobre los parámetros, como se muestra en la sección 4.2.

4.2.4.4. Datos con censura Tipo II

Considere un estudio donde las unidades ingresan al mismo tiempo y finaliza después de que r de las n unidades han experimentado el evento (o después de un tiempo fijo t_f). Los datos de tiempo ordenados son

$$t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(r)} = t^{+}_{(r+1)} = \cdots = t^{+}_{(n)}.$$

Si los tiempos siguen una distribución Weibull con función de densidad (3.4.1), los EMV de θ y α se obtienen resolviendo el sistema $U(\theta, \alpha) = 0$, este es:

$$\mathcal{R} - \theta^{\alpha}\left[\sum_{i=1}^{r} t^{\alpha} + (n-r)t_{(r)^{\alpha}}\right] = 0$$

$$\frac{\mathcal{R}}{\alpha} + \mathcal{R}\ln\theta + \sum_{i=1}^{r}\ln t_i + \theta^{\alpha}\left[\sum_{i=1}^{r} {t_i}^{\alpha}(\ln\theta + \ln t_i) + (n-r)t_i^{\alpha}(\ln\theta + \ln t_{(r)})\right] = 0. \quad (4.2.35)$$

Nuevamente, encontrar una solución al sistema de ecuaciones (4.2.35) puede no ser algebraicamente sencillo. Para ello se dispone de métodos numéricos incorporados en la mayoría de los paquetes estadísticos, como R, SAS, SPSS y MINITAB, entre otros, con los cuales se pueden obtener soluciones aproximadas a estas ecuaciones. Tales soluciones se toman como estimaciones de los parámetros para los datos disponibles.

4.3. Ajuste a los datos de un modelo probabilístico

Evidenciar y decidir sobre el modelo probabilístico que genera un conjunto de datos disponible es uno de los problemas transversal a la mayoría de las técnicas estadísticas.

Como se observa en la sección anterior, en la estimación de los parámetros de un modelo probabilístico se supone que los datos constituyen una muestra aleatoria de una distribución que, exceptuando sus parámetros, es conocida. De igual manera, para la aplicación de algunos de los métodos estadísticos de la variable aleatoria *tiempo hasta un evento* (sobrevida), es importante decidir y avalar si un conjunto de datos sigue la distribución sobre

la cual se erige el método, lo que aplica a modelos probabilísticos tales como el exponencial, Weibull, log-normal, log-logístico, gama, entre otros.

En esta sección se muestran algunas herramientas de tipo exploratorio y descriptivo como las gráficas cuantil-cuantil y las de riesgo. Con estas estrategias "visuales", se puede diagnosticar e hipotetizar el ajuste o no de un conjunto de datos a una distribución particular. No obstante, se dispone de algunas técnicas de diagnóstico de tipo confirmatorio acerca de la distribución probabilística que sigue un conjunto de datos, tales como la estadística ji-cuadrado, la razón de verosimilitud, el Criterio de Información de Akaike (AIC, por sus siglas en inglés) o el Criterio de Información de Bayes (BIC, por sus siglas en inglés). Estas técnicas también son útiles y están incorporadas en la mayoría de los paquetes para procesamiento estadístico.

En resumen, se procede de manera natural a como se trabaja en estadística, es decir, con base en un protocolo del estudio (experimental, cuasi-experimental u observacional), se obtiene un conjunto de datos sobre unas unidades de observación. De esta manera, el primer paso es la exploración para la descripción de los datos y un segundo paso es la abstracción sobre los modelos probabilísticos que más probablemente generan estos datos. En consecuencia, para el análisis de datos sobre tiempo para un evento, se explora la función empírica de sobrevida o la función de riesgo; el resultado de esta exploración debe advertir sobre el modelo más apropiado que sigue o se ajusta a los datos (y no al revés). Una primera estrategia exploratoria consiste en la comparación de la función de sobrevida del modelo propuesto con el estimador de Kaplan-Meier. En este procedimiento, se ajustan los modelos propuestos al conjunto de datos (por ejemplo, los modelos exponencial, log-normal, Weibull o gama, entre otros) y, a partir de las estimaciones de los parámetros de cada modelo, se estiman sus respectivas funciones de supervivencia, representadas por $S_{M_j}(t)$ para el j-ésimo modelo propuesto o asumido. Para el conjunto de datos, se obtiene la estimación, mediante Kaplan-Meier, $\tilde{S_{KM}}(t)$, para la función de sobrevida de los datos.

Luego, se comparan gráficamente las funciones de sobrevida estimadas para cada modelo propuesto, $\tilde{S}_{M_j}(t)$ con $\tilde{S_{KM}}(t)$. El modelo adecuado para los datos es aquel cuya curva de sobrevida esté uniformemente más cerca o próxima a la estimada mediante Kaplan-Meier; por *uniformemente* se entiende que lo está a través de todo el intervalo del estudio u observación.

En la figura 4.3 se observa que la función de sobrevida más cercana a la función de sobrevida estimada por el método de Kaplan-Meier, $\tilde{S}(t)_{KM}$, es la asociada con el modelo 1, $\tilde{S}(t)_{M_1}$; en consecuencia, se puede considerar el modelo probabilístico 1 (M_1) como el modelo relativamente más apropiado para estos datos. Este procedimiento pone de relieve que se trata de que el

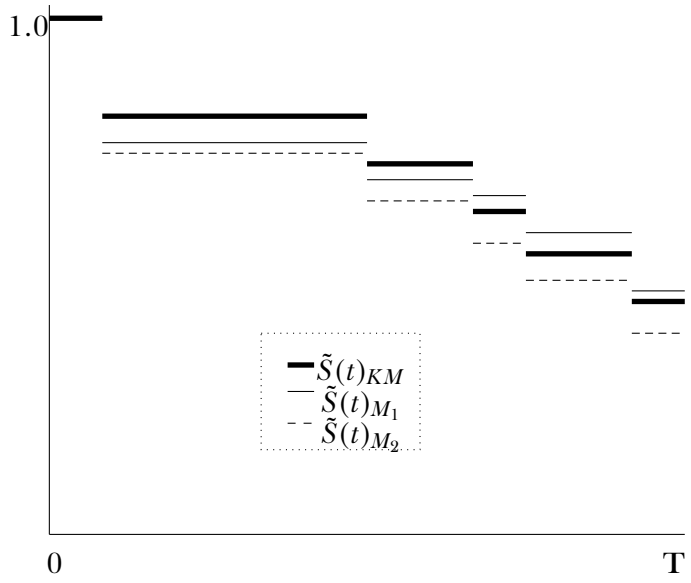

Figura 4.3. Escogencia de un modelo para los datos observados.

modelo siga a los datos y no al revés. Por eso se intentó con dos modelos, M_1 y M_2, para evidenciar cuál de los dos se ajusta (sigue) más a los datos, pues se trata de interpretar la realidad contenida en los datos y no de falsearla para sostener un modelo.

A continuación, se desarrolla esquemáticamente una segunda herramienta exploratoria para el diagnóstico del ajuste de un modelo probabilístico a un conjunto de datos.

Considere un conjunto de datos de tiempo registrados, $t_1, t_2, \ldots, t_n$. Nuevamente, se escriben los datos ordenados, subindexados entre paréntesis, para señalar el orden o puesto que ocupan frente a los demás datos. Así el ordenamiento del mínimo al máximo de los datos es

$$t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}.$$

Con esta notación, $t_{(1)} = \text{mín}\{t_1, t_2, \ldots, t_n\}$ es el tiempo mínimo hasta el evento, y $t_{(n)} = \text{máx}\{t_1, t_2, \ldots, t_n\}$, el tiempo máximo hasta el evento.

Un estimador de la función de sobrevida o confiabilidad, $S(t)$, es la función de sobrevida empírica (1.4.1), la cual se escribe nuevamente:

$$\tilde{S}_n(t) = \frac{No.obs > t}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{[t,\infty)}(T_i).$$

Para calificar la precisión de este estimador en cualquier punto $\mathbf{T} = t^*$, se estima un intervalo de confianza para la probabilidad de que el evento ocurra

después del tiempo t^*; el ancho del intervalo se relaciona de manera inversa con la precisión. Un intervalo de confianza de dos veces el error estándar es

$$\tilde{S}_n(t^*) \mp 2ee[\tilde{S}_n(t)]$$

$$\tilde{S}_n(t^*) \mp 2\sqrt{\frac{\tilde{S}_n(t^*)(1-\tilde{S}_n(t^*))}{n}}.$$

La representación gráfica de $\tilde{S}_n(t)$ respecto de t puede insinuar o advertir sobre la forma de la verdadera función de sobrevida $S(t)$. Como se muestra en la figura 4.4, $\tilde{S}_n(t)$ inicia en 1.0 para $t = 0$, pues a toda las unidades les ocurrirá el evento después de cero. Después, en la medida que t transcurre (incrementa), la función permanece constante hasta que ocurre el evento en una primera unidad (en el tiempo $t_{(1)}$); así, $\tilde{S}_n(t)$ reduce su altura en $\frac{1}{n}$. Nuevamente permanece constante en esta última altura hasta que a una segunda unidad le ha ocurrido el evento (en el tiempo $t_{(2)}$), entonces, la función de sobrevida reduce su altura en $\frac{1}{n}$ y así reiteradamente hasta que en $t = t_{(n)}$ la función alcanza el valor cero. Así, la función tiene una gráfica escalonada lineal con "saltos" hacia abajo de $\frac{1}{n}$ en cada uno de los puntos observados y ordenados en el tiempo. En caso de presentarse empates, por ejemplo, que k unidades alcancen el evento de manera simultánea, entonces el salto o caída será de $\frac{k}{n}$ en el momento de empate. La gráfica de $\tilde{S}_n(t)$ corresponde a una función escalonada no creciente y continua por la derecha (círculos "rellenos").

Mediante esta figura es difícil responder la pregunta: para un punto en el eje vertical, es decir, un valor de la sobrevida teórica (función $S(t)$), que naturalmente es un número en el intervalo $[0, 1]$, ¿cuál es el valor del tiempo que le corresponde en el eje horizontal? De otra manera, dada una fracción particular, ¿cuál de los datos (tiempo) tiene esta fracción de datos menor que sí misma? Para responder a estas preguntas se definen los *cuantiles muestrales*.

Para las definiciones y propiedades que se presentan a continuación, tenga en cuenta las siguientes propiedades:

- Si T es una variable aleatoria de tiempo hasta un evento con función de densidad de probabilidad $f(t)$, entonces la función de distribución acumulada de T , $F(t)$, es una función monótona creciente; es decir, tiene inversa, $F^{-1}(t)$.

- $U = F(t)$ es una variable aleatoria con distribución uniforme en el intervalo $[0, 1]$ (teorema de la transformación integral).

- La función de sobrevida $S(t) = 1 - F(t)$.

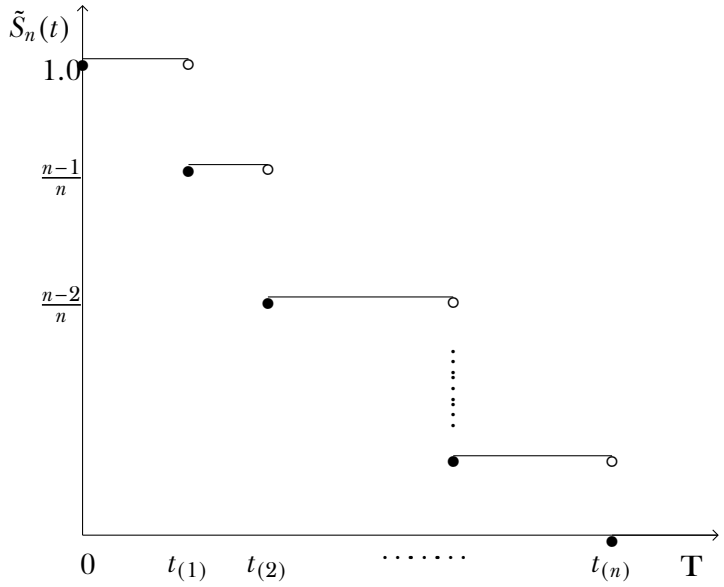

Figura 4.4. Función de sobrevida empírica.

Para un valor $0 \leq u \leq 1$, la *función cuantil muestral*, $Q_n(u)$, es

$$Q_n(u) = \text{ínf}\{t : \tilde{S}_n(t) \leq 1 - u\}. \tag{4.3.1}$$

La respuesta a la pregunta anterior implica la determinación del punto en el eje horizontal (tiempo) que se corresponde con el punto entre 0 y 1 seleccionado en el eje vertical. La anterior definición hace que Q_n sea la inversa de la función $1 - \tilde{S}_n$; es decir, $Q_n = \tilde{S}_n(1-u)$. La figura 4.5 ilustra esta construcción. Se toma cualquier valor u entre 0 y 1 y se busca el rango u orden j determinado por los datos, tal que $\frac{j-1}{n} < u < \frac{j}{n}$. Luego se traza $1-u$ una linea horizontal desde el punto $(0, 1-u)$ hasta la intersección con un "salto" de la función de sobrevida empírica $\tilde{S}_n(\cdot)$. Desde el punto de intersección se encuentra su proyección sobre el eje del tiempo; este punto de proyección corresponde a $Q_n(u) = t_{(j)}$. Convencionalmente, se selecciona el valor de u como el punto medio entre $\frac{j-1}{n}$ y $\frac{j}{n}$ para ser asociado con $t_{(j)}$. Este valor convencionalmente seleccionado está en p_j, donde

$$p_j = \frac{j - \frac{1}{2}}{n},$$

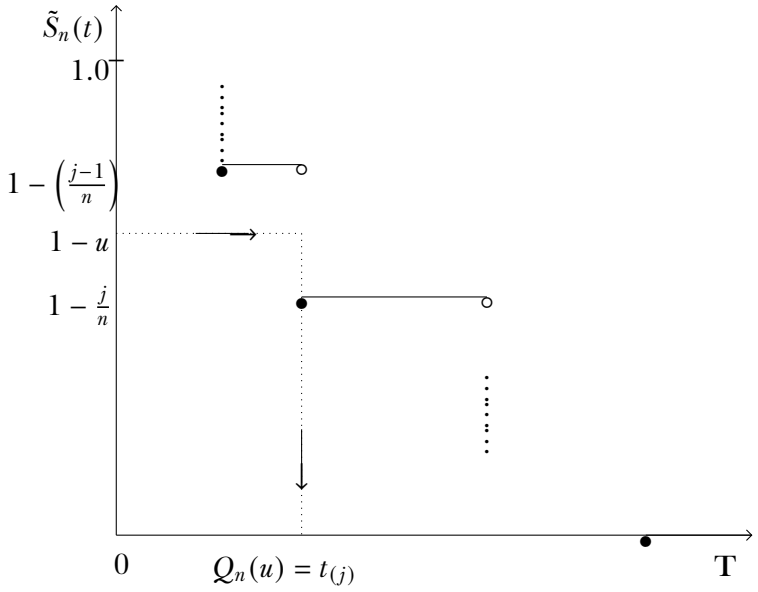

Figura 4.5. Función de sobrevida empírica y cuantil muestral.

se denomina el j-ésimo posición de punteo. Esto significa que $t_{(j)}$ es el $100\frac{j-\frac{1}{2}}{n}$-ésimo *percentil muestral* o el $\frac{j-\frac{1}{2}}{n}$-ésimo *cuantil muestral*, puesto que

$$Q_n(p_j) = t_{(j)}, \text{ para } p_j = \frac{j-\frac{1}{2}}{n}.$$

Se denomina un *gráfico de sobrevida empírico* al gráfico de posiciones de punteo frente a los datos ordenados; es decir, el gráfico está conformado por los pares $(t_{(j)}, 1-p_j)$, para $j = 1, 2, \ldots, n$.

El *p-ésimo cuantil poblacional*, ξ_p, es dado por la solución a la ecuación $p = F(\xi_p)$. Cuando ξ_p es escrito como una función $\xi_p = Q(p)$ de p, entonces Q es la T *función cuantil poblacional.*

Se tiene entonces que

$$\xi_p = Q(p) = F^{-1}(p) = S^{-1}(1-p),$$

por tanto Q_n es un estimador de Q.

Gráficos de probabilidad

Para variables continuas, un gráfico de probabilidad suministra evidencia de si un conjunto de datos procede de un modelo probabilístico específico.

Sobre la base de un modelo probabilístico S, la proporción de datos que el modelo pronostica como menores que el j-ésimo punto de los datos más grande debería estar cerca a p_j; es decir, la proporción de datos que excede al j-ésimo punto de datos más grande debería estar cerca a $1 - p_j$. Esto es,

$$F(t_{(j)}) \approx \frac{j - \frac{1}{2}}{n}, \text{ y } S(t_{(j)}) \approx \left(1 - \frac{j - \frac{1}{2}}{n}\right).$$

Un *gráfico o diagrama de probabilidad* tipo cuantil-cuantil (en inglés, *QQ-plot*) es un gráfico de los puntos

$$(Q(p_j), Q_n(p_j)) = \left(F^{-1}\left(\frac{j - \frac{1}{2}}{n}\right), t_{(j)}\right), \ j = 1, 2, \ldots, n,$$

los cuales se ubican en un sistema de coordenadas cartesiano.

Por la definición de diagrama de probabilidad, si los puntos se ubican cerca a una línea recta que contiene el origen $(0, 0)$ y tiene pendiente igual a 1 (inclinación de 45^o), entonces el modelo probabilístico se ajusta adecuadamente a los datos. A veces se requiere de transformaciones sobre uno o ambos ejes con el propósito de obtener una línea recta, de manera que el ajuste de los datos a una línea recta debe considerar la escala inducida por la transformación. También se puede presentar que, en los extremos (colas) de los datos ordenados, estos queden por encima o por debajo de la línea recta, mientras que la mayoría de los datos intermedios se aproximan a la línea recta (modelo propuesto); es decir, la nube de puntos tiene la apariencia de una letra "S" alargada, como el símbolo de la integral. Esta última característica, si bien no hace que se descarte el modelo, debe hacer que se revisen los datos de los extremos dentro del marco conceptual de estos y la manera como fueron obtenidos, pues pueden haber situaciones de adaptación, fatiga, descalibración de instrumentos, etc.

4.3.0.1. Gráficos de probabilidad para la distribución exponencial

Se quiere inspeccionar gráficamente si la variable aleatoria *tiempo hasta la ocurrencia de un evento*, $\mathbf{T}$, tiene distribución exponencial, $\exp(\theta)$. Su función de sobrevida es

$$S(t) = \exp(-\theta t), \ t > 0. \tag{4.3.2}$$

Para la construcción del gráfico de probabilidad se reemplaza t por $t_{(j)}$ en (4.3.2). Luego, tomando logaritmos en los dos miembros de la igualdad

$$\ln S(t_{(j)}) = -\left(\theta t_{(j)}\right), \ j = 1, 2 \ldots, n,$$

desde esta expresión se obtiene

$$t_{(j)} = \frac{1}{\theta}[-\ln S(t_{(j)})],\ j = 1, 2 \ldots, n, \tag{4.3.3}$$

esta es una relación lineal entre $t_{(j)}$ y $[-\ln S(t_{(j)})]$.

A partir de la expresión (4.3.3), se tiene que si $[-\ln S(t_{(j)})] = 1$, entonces $t_{(j)} = 1/\theta$. Este resultado puede emplearse para estimar tanto $1/\theta$ como θ desde la recta ajustada. Así, el valor de $t_{(j)}$ en $[-\ln S(t_{(j)})] = 1$ es un estimador de la media de la distribución exponencial, $1/\theta$ y su inversa es un estimador de la tasa de riesgo θ.

Si una distribución exponencial se ajusta a los datos, entonces

$$S(t_{(j)}) \approx 1 - \left(\frac{j - \frac{1}{2}}{n}\right),\ j = 1, 2 \ldots, n.$$

Si se reemplaza esta aproximación (valor empírico) en (4.3.3), entonces, cuando el modelo se ajuste adecuadamente a los datos, se debe tener que

$$t_{(j)} \approx \frac{1}{\theta}\left[-\ln\left(1 - \left(\frac{j - \frac{1}{2}}{n}\right)\right)\right]. \tag{4.3.4}$$

Lo anterior significa que para una distribución exponencial, el gráfico o nube de puntos

$$\left(-\ln\left[1 - \left(\frac{j - \frac{1}{2}}{n}\right)\right], t_{(j)}\right) \tag{4.3.5}$$

debería aproximarse a una línea recta con pendiente $\frac{1}{\theta}$. Si el ajuste a una línea recta es bueno, entonces el valor de la pendiente observada en la recta de "mejor" ajuste es aproximadamente $\frac{1}{\theta}$, cantidad cuyo inverso se puede considerar como una estimación de θ, la tasa de riesgo.

Ejemplo 4.3.1. Suponga que 21 pacientes con leucemia aguda tienen los siguientes tiempos (en meses) de remisión: 1, 1, 2, 2, 3, 4, 4, 5, 5, 6, 8, 8, 9, 10, 10 12, 14, 16, 20, 24 y 34. Estos datos fueron tomados de Lee y Wang [44, pp. 204-206]. Se quiere saber si los datos de remisión proceden de una distribución exponencial. La tabla 4.4 contiene los tiempos de remisión ordenados, $t_{(j)}$, los valores de la función de sobrevida empírica en cada remisión, $\tilde{S}_n(\cdot)$, y su logaritmo con signo opuesto, $-\ln \tilde{S}_n(\cdot)$; es decir, los cuantiles muestrales de acuerdo con los datos disponibles. En este caso hay presencia de datos empatados, para los cuales se calcula la función de sobrevida empírica para solo uno de ellos; posteriormente, se hace un tratamiento estadístico para datos empatados.

Tabla 4.3. Remisión en estudio de leucemia aguda

Orden j	$t_{(j)}$	$\tilde{S}_n(\cdot) \approx \left(1 - \frac{j-\frac{1}{2}}{21}\right)$	$-\ln \tilde{S}_n(\cdot) \approx -\ln\left[1 - \left(\frac{j-\frac{1}{2}}{21}\right)\right]$
1	1		
2	1	0.9286	0.0741
3	2		
4	2	0.8333	0.1823
5	3	0.7857	0.2412
6	4		
7	4	0.6905	0.3704
8	5		
9	5	0.5952	0.5188
10	6	0.5476	0.6022
11	8		
12	8	0.4524	0.7932
13	9	0.4048	0.9045
14	10		
15	10	0.3095	1.1727
16	12	0.2619	1.3398
17	14	0.2143	1.5404
18	16	0.1667	1.7918
19	20	0.1190	2.1282
20	24	0.0714	2.6391
21	34	0.0238	3.7377

En el gráfico de probabilidad que se muestra en la figura 4.6 se ajusta una línea recta. A simple vista, se observa que el modelo exponencial se ajusta adecuadamente a los datos (nuevamente, el modelo sigue a los datos). El punto donde $[-\ln S(t_{(j)})] = 1$, desde la figura 4.6, es aproximadamente $t_{(j)} = 9.0$ meses. Este es un estimador de la media $1/\theta$, por lo que un estimador de la tasa de riesgo es $\hat{\theta} = 1/9 = 0.111$ por mes. El EMV de θ, de acuerdo con la expresión (4.2.21), es $\hat{\theta} = 21/198 = 0.107$, el cual es muy próximo al obtenido gráficamente.

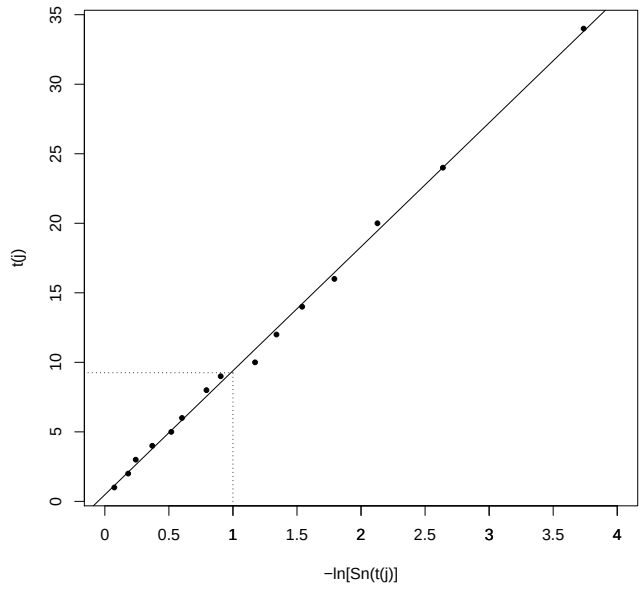

Figura 4.6. Gráfico de probabilidad: ajuste de distribución exponencial.

4.3.0.2. Gráficos de probabilidad para la distribución Weibull

La función de sobrevida o de confiabilidad de una variable aleatoria **T** con distribución $Weibull(\theta, \alpha)$ es

$$S(t) = 1 - F(t) = e^{-(\theta t)^{\alpha}},\ \theta,\ \alpha\ y\ t \geq 0.$$

El gráfico de probabilidad se construye reemplazando t por $t_{(j)}$ en la función de sobrevida anterior. Aplicando logaritmos en los dos miembros de la igualdad, esta se transforma en

$$\ln S(t_{(j)}) = -(\theta t)^{\alpha}.$$

Una vez se aplican logaritmos en los dos miembros de la igualdad, este logaritmo puede ser el de base e o el de base 10. La diferencia está en la escala, pues el logaritmo en base 10 "acerca" más los valores grandes que el logaritmo natural. En este texto se aplica el logaritmo natural, el cual conlleva al siguiente resultado:

$$\ln[-\ln S(t_{(j)})] = \alpha[\ln t_{(j)} + \ln \theta].$$

A partir de la última expresión, se identifica (despejando adecuadamente) la línea recta,

$$\ln t_{(j)} = \ln \frac{1}{\theta} + \frac{1}{\alpha} \ln[-\ln S(t_{(j)})]. \tag{4.3.6}$$

Si el modelo probabilístico de Weibull se ajusta a los datos adecuadamente, entonces $S(t_{(j)}) \approx 1 - \left(\frac{j-\frac{1}{2}}{n}\right)$. Al sustituir esta aproximación de la función de sobrevida en la recta de ecuación (4.3.6), si el ajuste del modelo a los datos es adecuado, entonces

$$\ln t_{(j)} \approx \ln \frac{1}{\theta} + \frac{1}{\alpha} \ln \left[-\ln \left(1 - \frac{j-\frac{1}{2}}{n} \right) \right]. \tag{4.3.7}$$

Así, el gráfico de probabilidad se conforma por los puntos de coordenadas

$$\left(\ln \left[-\ln \left(1 - \frac{j-\frac{1}{2}}{n} \right) \right], \ \ln t_{(j)} \right), \ j = 1, 2 \ldots, n. \tag{4.3.8}$$

Si la nube conformada por estos puntos se ajusta adecuadamente a una línea recta, entonces la distribución de probabilidad subyacente es de tipo Weibull. La pendiente de la recta es $\frac{1}{\alpha}$ y el intersecto es $\ln \theta$; si $\theta = \alpha = 1$, se tiene una línea recta a 45° (bisectriz principal).

Ejemplo 4.3.2. Veinte motores de refrigerador se pusieron en funcionamiento, bajo condiciones extremas, hasta que se fundieron. Los tiempos hasta la falla (en horas) se registraron como se muestra en Chatfield [9, p. 16]. Se trata de verificar si la distribución de Weibull sigue a los datos, es decir, si estos datos se generan desde un modelo probabilístico Weibull. Los tiempos hasta la falla ordenados, $t_{(j)}$, la función de sobrevida empírica evaluada en cada punto del tiempo de fundición de los 20 motores y los demás términos componentes de la ecuación de la línea recta expresada en (4.3.7) están contenidos en la tabla 4.4. Los puntos del tiempo hasta la fundición de los $n = 20$ motores, calculados mediante (4.3.8), se ubican e ilustran en el gráfico de probabilidad 4.7. En el gráfico se ajusta, de manera empírica ("a ojo") una línea recta. Se observa que los puntos están ubicados en torno o sobre esta línea recta a una distancia muy pequeña y, para algunos, nula. Esta característica sugiere que el modelo Weibull se ajusta razonablemente a estos datos.

El parámetro de forma α se puede estimar gráficamente como el inverso de la pendiente de la línea recta ajustada y trazada en la gráfica de probabilidad. Si el ajuste de estos puntos a una línea recta es apropiado, entonces en el valor 0 sobre el eje horizontal, es decir, donde $\ln[-\ln S(t_{(j)})] = 0$, el correspondiente en el eje vertical $\ln t_{(j)}$ es un estimador (por (4.3.6)) de

Tabla 4.4. Fundición de motores de refrigeración

$t_{(j)}$	j	$\frac{j-\frac{1}{2}}{n}$	$\ln\left[-\ln\left(1-\frac{j-\frac{1}{2}}{n}\right)\right]$	$\ln t_{(j)}$
104.3	1	0.025	-3.676	4.647
158.7	2	0.075	-2.552	5.067
193.7	3	0.125	-2.013	5.266
201.3	4	0.175	-1.648	5.305
206.2	5	0.225	-1.367	5.329
227.8	6	0.275	-1.134	5.428
249.1	7	0.325	-0.934	5.518
307.8	8	0.375	-0.755	5.729
311.5	9	0.425	-0.592	5.741
329.6	10	0.475	-0.440	5.798
358.5	11	0.525	-0.295	5.882
364.3	12	0.575	-0.156	5.898
370.4	13	0.625	-0.019	5.915
380.5	14	0.675	0.117	5.941
394.6	15	0.725	0.255	5.978
426.2	16	0.775	0.400	6.055
434.1	17	0.825	0.556	6.073
552.6	18	0.875	0.732	6.315
594.0	19	0.925	0.952	6.387
691.5	20	0.975	1.305	6.539

$\ln(1/\theta)$); se tiene así un estimador, desde el gráfico de probabilidad Weibull, de $1/\theta$ y, por lo tanto, de θ.

A partir de la recta que sigue los puntos (modelo que sigue los datos), para el valor $\ln[-\ln S(t_{(j)})] = 0$ en el eje horizontal, el valor correspondiente sobre el eje vertical es $\ln t_{(j)} \approx 5.9$. Ahora, desde la ecuación (4.3.6), se tiene:

$$\text{como } \ln t_{(j)} = \ln \frac{1}{\theta}, \text{ entonces } 5.9 \approx \ln \frac{1}{\tilde{\theta}}, \text{ así, } \frac{1}{\tilde{\theta}} \approx \exp(5.9);\ \tilde{\theta} \approx 0.003.$$

Se escribe la "estimación" empírica $\tilde{\theta}$ no solo por la ubicación gráfica (estimación "a ojo") sino también por la aproximación.

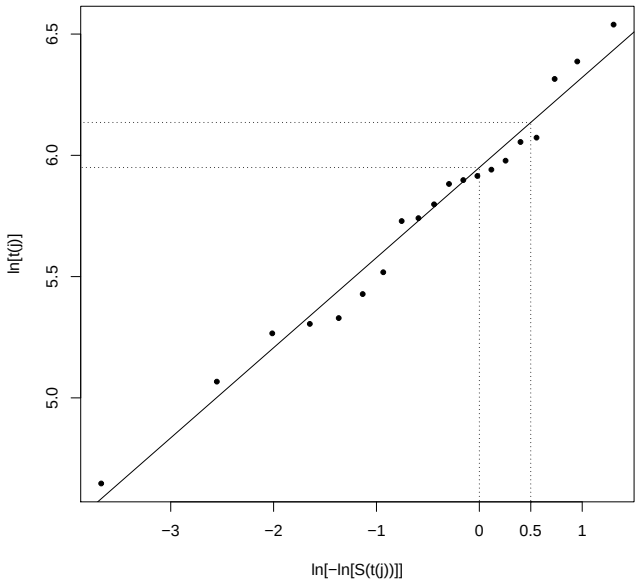

Figura 4.7. Gráfico de probabilidad: ajuste de distribución Weibull.

Ahora, en $\ln[-\ln S(t_{(j)})] = 0.5$ el valor de $\ln t_{(j)} \approx 6.1$, por lo que, desde la ecuación (4.3.6), se tiene:

$$\ln t_{(j)} = \ln \frac{1}{\theta} + \frac{0.5}{\alpha}, \text{ con } \tilde{\theta} \approx 0.003 \text{ y } \ln t_{(j)} \approx 6.1, \text{ entonces } \tilde{\alpha} \approx 1.72.$$

En resumen, se puede afirmar que el modelo que se ajusta moderadamente bien a estos datos es un modelo *Weibull* (0.003, 1.72); de otra manera, se puede suponer que los datos proceden de un modelo $Weibull(0.003, 1.72)$.

4.3.0.3. Gráficos de probabilidad para la distribución log-normal

La variable aleatoria $\mathbf{T}$ tiene una distribución log-normal con parámetros μ y σ^2, entonces, $\ln \mathbf{T}$ tiene distribución normal con media μ y varianza σ^2. La conocida "estandarización" de la variable aleatoria $\ln \mathbf{T}$ es $Z = (\ln \mathbf{T} - \mu)/\sigma$, la cual tiene distribución normal estándar y se nota por $\phi(z)$. La distribución log-normal acumulada es

$$F(t) = \Phi\left(\frac{\ln t - \mu}{\sigma}\right), \ t > 0,$$

donde $\Phi(\cdot)$ es la distribución acumulada normal estándar, con μ y σ la media y la desviación estándar de $\ln \mathbf{T}$.

Una gráfica de probabilidad para la distribución log-normal se basa en la siguiente ecuación:

$$\ln t = \mu + \sigma \Phi^{-1}\left(F(t)\right), \tag{4.3.9}$$

la cual corresponde a una línea recta con pendiente σ e intersecto μ.

A partir de la ecuación (4.3.9), si $\Phi^{-1}\left(F(t)\right) = 0$, entonces $\ln t = \mu$. De manera similar, si $\Phi^{-1}\left(F(t)\right) = 1$, entonces nuevamente, a partir de la ecuación (4.3.9), $\sigma = \ln t - \mu$.

Si el modelo probabilístico log-normal se ajusta razonablemente bien a los datos, entonces $F(t_{(j)}) \approx \left(\frac{j-\frac{1}{2}}{n}\right)$. Al sustituir esta aproximación de la función de distribución acumulada en la recta de ecuación (4.3.9), si el ajuste del modelo a los datos es apropiado, entonces

$$\ln t_{(j)} \approx \mu + \sigma \Phi^{-1}\left(\frac{j-\frac{1}{2}}{n}\right), \ j = 1, 2 \ldots, n. \tag{4.3.10}$$

La gráfica de probabilidad se construye con los puntos de coordenadas

$$\left(\Phi^{-1}\left(\frac{j-\frac{1}{2}}{n}\right), \ \ln t_{(j)}\right), \ j = 1, 2 \ldots, n. \tag{4.3.11}$$

Si existe una línea recta tal que el conjunto de puntos (nube), cuyas coordenadas se calculan mediante (4.3.11), se dispone muy cerca de la línea recta, se puede advertir que el modelo log-normal es apropiado para tales datos.

Ejemplo 4.3.3. Suponga que a un nuevo tipo de batería para teléfono móvil se le registra su tiempo de duración (en meses). Para ello se seleccionaron aleatoriamente 15 de estas baterías de un lote de producción, y se hizo seguimiento del tiempo de vida útil en cada una. Los tiempos fueron: 3, 5, 6, 7, 8, 10, 11, 13, 15, 16, 21, 24, 43, 46 y 55.

En consecuencia, a partir de la figura 4.8, la media, μ, se estima por el intersecto de la recta con el eje vertical y la desviación estándar, σ, se estima como la pendiente de la recta ajustada en la gráfica de probabilidad. De manera más explícita, tomando sobre el eje horizontal el punto donde $\Phi^{-1}\left((j-0.5)/n\right) = 0$, se encuentra el valor correspondiente en el eje vertical, valor con el cual se tiene una estimación empírica (gráfica) de μ; para estos datos, $\tilde{\mu} \approx 2.6$. Ahora, de manera similar, en el eje horizontal se ubica el punto donde $\Phi^{-1}\left((j-0.5)/n\right) = 1$, por lo que sobre la gráfica ("a ojo") se ubica el valor en el eje vertical correspondiente. Este es aproximadamente $\ln t_{(j)} \approx 3.4$, entonces, $\tilde{\sigma} \approx 3.4 - 2.6$, es decir $\tilde{\sigma} \approx 0.8$.

Tabla 4.5. Tiempo a la falla de baterías de teléfono móvil

j	$t_{(j)}$	$\frac{j-\frac{1}{2}}{n}$	$\Phi^{-1}\left(\frac{j-\frac{1}{2}}{n}\right)$	$\ln t_{(j)}$
1	3	0,033	-1,834	1,099
2	5	0,100	-1,282	1,609
3	6	0,167	-0,967	1,792
4	7	0,233	-0,728	1,946
5	8	0,300	-0,524	2,079
6	10	0,367	-0,341	2,303
7	11	0,433	-0,168	2,398
8	13	0,500	0,000	2,565
9	15	0,567	0,168	2,708
10	16	0,633	0,341	2,773
11	21	0,700	0,524	3,045
12	24	0,767	0,728	3,178
13	43	0,833	0,967	3,761
14	46	0,900	1,282	3,829
15	55	0,967	1,834	4,007

Datos simulados por el autor.

A partir de este ejercicio exploratorio, se puede afirmar que hay evidencia empírica para sostener que los datos son generados por un modelo *log-normal* (2.6, 0.8).

4.3.0.4. Gráfica de la función de riesgo

Una gráfico de la función de riesgo es similar a la gráfica de probabilidad. La principal diferencia está en que el tiempo para el evento, o una función de este, es graficado frente a la función de riesgo acumulada (una función de esta). La gráfica de riesgo se adapta para la consideración e inclusión de datos censurados. Tal como en las gráficas de probabilidad, desde las gráficas de riesgo se pueden obtener estimaciones empíricas de los parámetros haciendo uso del mismo soporte computacional con el que se construyen las gráficas, tal como lo señalan Lee y Wang [44].

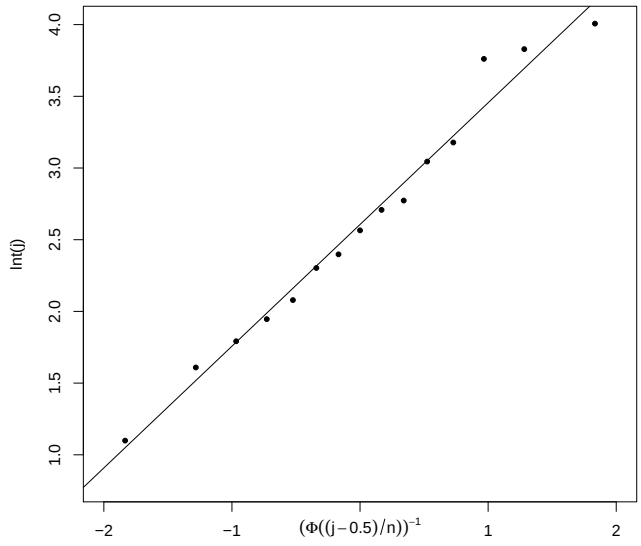

Figura 4.8. Gráfico de probabilidad: ajuste de distribución log-normal.

Para determinar si un conjunto de datos con observaciones censuradas procede de un modelo de probabilidad específico, se construye una gráfica de riesgo ubicando el tiempo hasta el evento (o una función de este) frente a la función de riesgo acumulado (o una función de esta). La función de riesgo acumulado se puede estimar mediante el siguiente procedimiento.

1. Ordene de manera creciente las observaciones de tiempo hasta el evento, considerando si algunas son censuradas. Para las observaciones empatadas (que tienen el mismo valor), sean censuradas o no, se les asigna el orden aleatorio tomado del conjunto de órdenes que tendrían si no estuvieran empatadas. Se debe asignar un rótulo o función indicadora para las observaciones censuradas y las no censuradas.

2. A las observaciones ordenadas de manera creciente hay que asignarles, a cada una, el orden inverso. Así, a la menor se le asignará el número n, a la segunda el número $n - 1$, y así reiteradamente hasta que a la mayor se le asigna el número 1. Los números asignados de esta forma se denominan valores K o números de orden inverso. Para las observaciones no censuradas, K es el número de unidades en riesgo en el respectivo momento.

3. Calcular el correspondiente riesgo para cada observación no censurada. En las observaciones censuradas no se evalúa el riesgo. El valor del riesgo en cada observación no censurada es $1/K$. Esta es la fracción de riesgo, entre K unidades a las cuales no les ha ocurrido el evento, de que este ocurra inmediatamente después. Esta es una probabilidad condicional empírica de que ocurra el evento entre las observaciones no censuradas.

4. Para cada una de las observaciones no censuradas, hay que calcular el valor del riesgo acumulado. Este corresponde a la suma de los riesgos individuales de las observaciones no censuradas que las preceden cronológicamente. Para las observaciones empatadas, el riesgo acumulado se evalúa únicamente en el menor K entre las observaciones no censuradas.

Sobre los 21 pacientes con leucemia aguda del ejemplo 4.3.0.1, se construye la tabla 4.6 siguiendo los pasos o el algoritmo descrito anteriormente. El propósito es verificar cuál modelo tiene mejor adaptación a estos datos. A continuación, se muestra la adaptación a estos datos del modelo de riesgo exponencial.

4.3.0.5. Modelo de riesgo exponencial

Para el modelo exponencial de la variable aleatoria *tiempo hasta el evento*, la función de riesgo y riesgo acumulado son, respectivamente,

$$h(t) = \theta \text{ y } H(t) = \theta t, \text{ con } \theta > 0 \text{ y } t > 0.$$

Así, el tiempo t puede escribirse como una función lineal en $H(t)$, es decir,

$$t_{(j)} = \frac{1}{\theta} H(t_{(j)}). \tag{4.3.12}$$

La pendiente de esta recta es la media de una distribución exponencial, es decir, a partir de la ecuación (4.3.12) es $1/\theta$ y ocurre en un valor de t para el cual $H(t) = 1$.

Para encontrar gráficamente un estimador de la media de remisión por leucemia aguda, se toma el valor de $\tilde{H}(t) = 0.5$, pues 1 está fuera del rango en la gráfica 4.9. Para $\tilde{H}(t_{(j)}) = 0.5$, $t_{(j)} = 1.69$, por lo que un estimador empírico de θ, por la ecuación (4.3.12), es $\tilde{\theta} = 0.5/1.69 = 0.0296$. Así, un estimador de la media del tiempo de remisión es aproximadamente 34 semanas.

Tabla 4.6. Tiempo a la muerte de pacientes con leucemia aguda

Tiempo $t_{(j)}$	Orden inverso: K	Riesgo empírico: $1/K$	Riesgo acumulado: $\tilde{H}(t)$
6	21	0.048	
6^+	20		
6	19	0.053	
6	18	0.056	0.156
7	17	0.058	0.215
9^+	16		
10	15	0.067	0.281
10^+	14		
11^+	13		
13	12	0.083	0.365
16	11	0.091	0.456
17^+	10		
19^+	9		
20^+	8		
22	7	0.143	0.598
23	6	0.167	0.765
25^+	5		
32^+	4		
32^+	3		
34^+	2		
35^+	1		

Tomado de Lee y Wang [44, p. 29].

4.3.0.6. Modelo de riesgo Weibull

Si la variable aleatoria *tiempo para el evento* tiene distribución de $Weibull(\theta, \alpha)$, su función de riesgo es

$$h(t) = \theta\alpha(\theta t)^{\alpha-1}, \text{ para } t > 0.$$

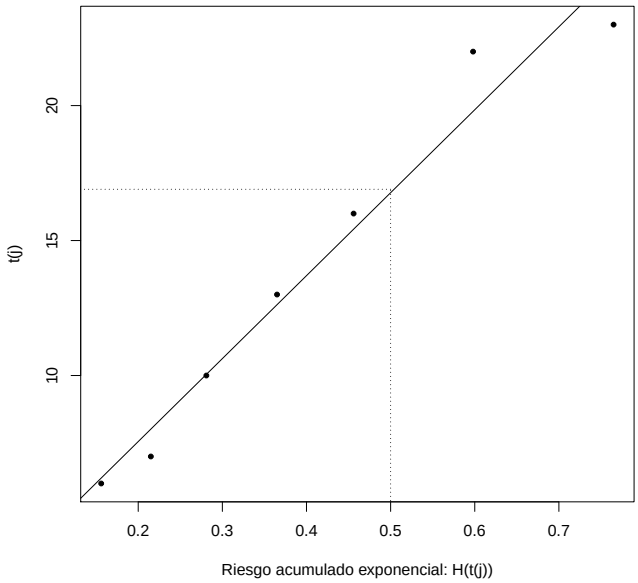

Figura 4.9. Gráfico de riesgo: ajuste de la distribución exponencial.

La función de riesgo acumulado es

$$H(t) = (\theta t)^{\alpha}, \text{ para } t > 0. \tag{4.3.13}$$

A partir de la ecuación (4.3.11), elevando a la potencia $1/\alpha$, tomando logaritmos en los dos miembros y luego de despejar $\ln t$, para los tiempos ordenados, $t_{(j)}$, se tiene

$$\ln t_{(j)} = \ln\left(\frac{1}{\theta}\right) + \frac{1}{\alpha}\ln H(t_{(j)}). \tag{4.3.14}$$

A partir de la gráfica de riesgo contenido en la figura 4.10, con la recta ajustada para estos datos, se observa que para $\ln H(t_{(j)}) = -1$, el valor $\ln t_{(j)} \approx 2.55$, mientras que para $\ln H(t_{(j)}) = -0.5$, el valor $\ln t_{(j)} \approx 3.01$ (valores encontrados "a ojo") sobre la gráfica de riesgo. De esta manera, por la ecuación (4.3.14), se tiene el sistema de ecuaciones

$$\begin{cases} 2.57 = \ln\left(\frac{1}{\theta}\right) - \frac{1}{\alpha} \\ 3.01 = \ln\left(\frac{1}{\theta}\right) - \frac{0.5}{\alpha}. \end{cases}$$

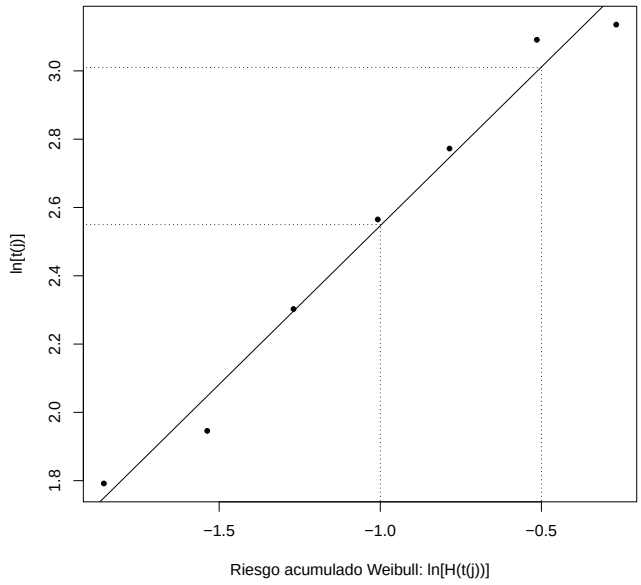

Figura 4.10. Gráfico de riesgo: ajuste de la distribución Weibull.

La solución de este sistema suministra una estimación empírica de los parámetros; estas son $\tilde{\theta} \approx 0.032$ y $\tilde{\alpha} \approx 1.14$. Así, se podría afirmar que el modelo $Weibull(0.032)$ sigue moderadamente los datos.

4.3.0.7. Modelo de riesgo log-normal

Su función de densidad log-normal está dada por:

$$f(t) = \frac{1}{t\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{\ln(t)-\mu}{\sigma}\right)^2\right]$$
$$= \frac{\phi\left(\frac{\ln(t)-\mu}{\sigma}\right)}{t\sigma}, \text{ con } -\infty < \mu < \infty,\ 0 < \sigma < \infty,$$

donde $\phi(\cdot)$ es la función de densidad de una variable aleatoria con distribución normal estándar.

La función de sobrevida está dada por:

$$S(t) = 1 - \Phi\left(\frac{\ln(t)-\mu}{\sigma}\right) = \Phi\left(\frac{-\ln(t)+\mu}{\sigma}\right),$$

donde Φ es la función de distribución acumulada de una variable aleatoria con distribución normal estándar.

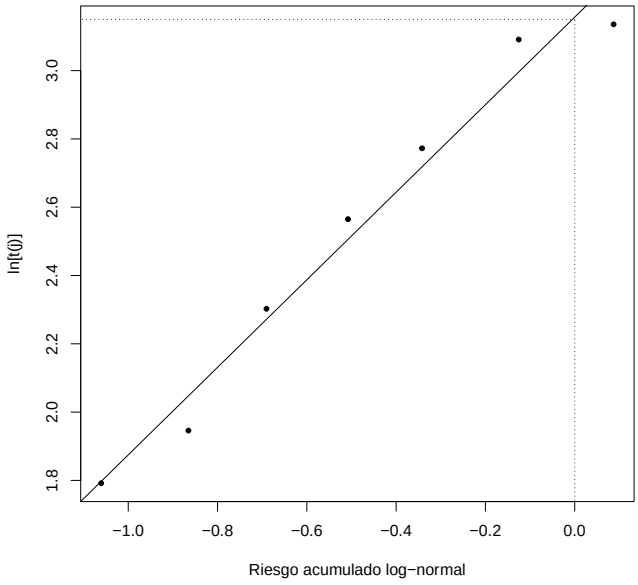

Figura 4.11. Función de sobrevida log-normal.

La función de riesgo de la distribución log–normal es

$$h(t) = \frac{\frac{\phi\left(\frac{\ln(t)-\mu}{\sigma}\right)}{t}}{1-\Phi\left(\frac{\ln(t)-\mu}{\sigma}\right)}.$$

La función de riesgo acumulada, $H(t)$, se obtiene como la integral en $[0, t]$ de la función de riesgo; esta es

$$H(t) = -\ln\left[1-\Phi\left(\frac{\ln t-\mu}{\sigma}\right)\right]. \tag{4.3.15}$$

A partir de (4.3.15), el logaritmo del tiempo ordenado para el evento, $\ln t_{(j)}$, es una función del riesgo acumulado, $H(t_{(j)})$, esta es

$$\ln t_{(j)} = \mu + \sigma\Phi^{-1}\left(1-e^{-H(t_{(j)})}\right), \tag{4.3.16}$$

donde $\Phi^{-1}(\cdot)$ es la inversa de la función de distribución acumulada de una normal estándar.

El logaritmo del tiempo al evento es una función lineal de $\Phi^{-1}\left(1-e^{-H(t)}\right)$. La gráfica de riesgo log-normal es una gráfica de $\Phi^{-1}\left(1-e^{-H(t)}\right)$ frente a $\ln t$, la cual, para un conjunto de datos específico, debe ser tal que si el modelo probabilístico log-normal sigue los datos, entonces la gráfica de los puntos (datos) debe ajustarse a una línea recta con pendiente igual a σ e intersecto igual a μ.

A partir de la gráfica 4.11 se observa que, para $\Phi^{-1}\left(1-e^{-H(t_{(j)})}\right)=0$, el valor correspondiente en el eje vertical es $\ln t_{(j)} \approx 3.2$. Entonces, por la ecuación de la línea recta ajustada (4.3.16), $\ln t_{(j)} = \mu$, por lo que $\tilde{\mu} \approx 25$.

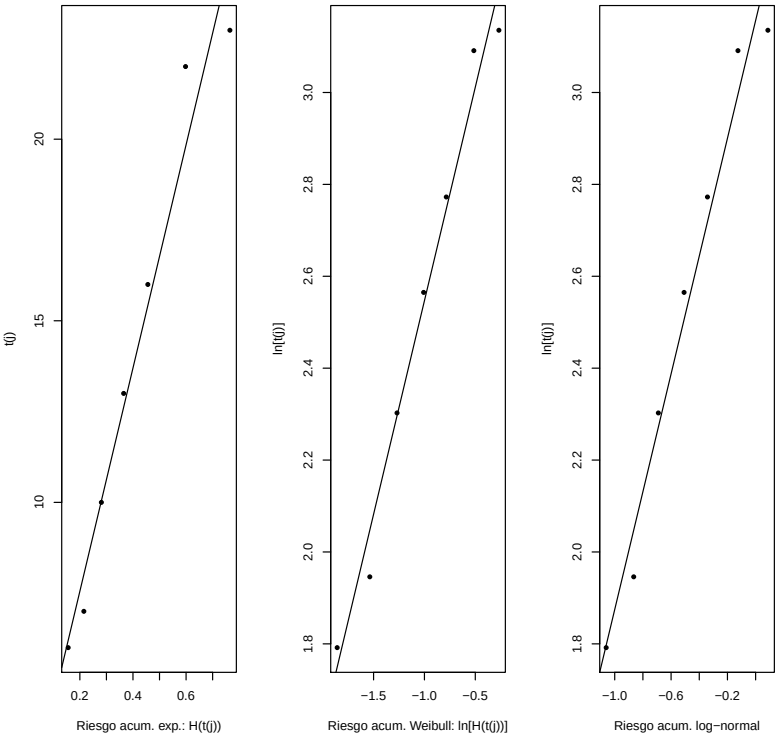

Figura 4.12. Gráficas de la función de riesgo.

En resumen, se puede concluir que de estos modelos el que mejor sigue a los datos es el modelo Weibull. La gráfica 4.12 contiene los diagramas de riesgo acumulado con ajuste lineal para los modelos exponencial, Weibull y log-normal. Por apreciación visual, se corrobora que el modelo más ajustado a estos datos es el Weibull.

4.3.0.8. Residuos de Cox-Snell

El método de los residuos de Cox-Snell se puede aplicar para diagnosticar la calidad de cualquier modelo paramétrico frente a un conjunto de datos. El residuo r_i de la i-ésima unidad con tiempo hasta el evento t_i, es definida como

$$r_{CSi} = -\ln \hat{S}(t_j), \text{ para } i = 1, 2, \ldots, n, \tag{4.3.17}$$

donde $\hat{S}(t)$ es la función de sobrevida basada sobre los estimadores de máxima verosimilitud de sus parámetros. El residual es censurado, o no, dependiendo de si t_i es o no es censurado. Además, por (1.5.4), la función de riesgo acumulado $H(t) = -\ln S(t)$, así que el residual de Cox-Snell es una estimación del riesgo acumulado en t_i. Una definición equivalente de los residuos de Cox-Snell es

$$r_{CSi} = \hat{H}(t_i), \text{ para } i = 1, 2, \ldots, n. \tag{4.3.18}$$

Los residuales de Cox-Snell tienen la propiedad de que si el modelo seleccionado se ajusta a los datos, entonces los residuales tienen distribución exponencial unitaria con función de densidad $f_R(r) = e^{-r}$.

Sea $S_R(t)$ la función de sobrevida de los residuales de Cox-Snell. Entonces,

$$S_R(r) = \int_r^\infty f_R(u)du = \int_r^\infty e^{-u}du0e^{-r}, \text{ y } -\ln S_R(r) = -\ln(e^{-r}) = r. \tag{4.3.19}$$

Los residuos de Cox-Snell para los modelos de regresión exponencial, Weibull y log-normal, respectivamente, son:

$$\begin{aligned} Exponencial: &\quad r_{CSi} = \left(t_i \exp[-\mathbf{x}_i'\beta]\right) \\ Weibull: &\quad r_{CSi} = \left(t_i \exp[-\mathbf{x}_i'\beta]\right)^{\hat{\alpha}} \\ log-normal: &\quad r_{CSi} = -\ln\left[1 - \Phi\left(\frac{\ln(t_i) - \mathbf{x}_i'\beta}{\hat{\sigma}}\right)\right]. \end{aligned} \tag{4.3.20}$$

Sea $\hat{S}_R(t)$ el estimador de Kaplan-Meier de $S_R(t)$. A partir de la expresión (4.3.19), es evidente que una gráfica de r_{CSi} frente a $-\ln \hat{S}_R(t)$ debería ser una línea recta con pendiente igual a la unidad e intersecto igual a cero, si la distribución de sobrevida es apropiada y sin importar la forma de la distribución. El procedimiento para usar los residuos de Cox-Snell se puede resumir en los siguientes pasos.

1. Encuentre los estimadores de máxima verosimilitud para los parámetros del modelo teórico seleccionado.

2. Calcule los residuales de Cox-Snell $r_{CSi} = -\ln \hat{S}(t_i)$, $i = 1, 2, \ldots, n$, donde $\hat{S}(t_i)$ es la función de sobrevida estimada a partir de los EMV de sus parámetros.

3. Estime, mediante el estimador de Kaplan-Meier, la función de sobrevida, $S_R(r)$, de los residuos r_{CSi} calculados en el paso 2. Luego, mediante $\hat{S}_R(r)$ calcule $-\ln \hat{S}_R(r_{CSi})$, para $i = 1, 2, \ldots, n$.

4. Grafique r_{CSi} frente a $-\ln \hat{S}_R(r_{CSi})$, $i = 1, 2, \ldots, n$. Si la nube de puntos definidos por las parejas $(r_{CSi}, -\ln \hat{S}_R(r_{CSi}))$ es una línea recta con pendiente igual a la unidad e intersecto cero, la distribución se ajusta adecuadamente a los datos.

   Si hay censura a derecha en algún tiempo para el evento t_i y el modelo asumido para los datos se ajusta adecuadamente a estos datos, el correspondiente residuo de Cox-Snell es $-\ln S(t^+) = H(t^+)$ y es menor que el residuo evaluado en una observación no censurada con el mismo valor t_i, dado que $H(t)$ es una función monótona creciente en t. Una de las modificaciones propuestas es tomar:

$$r'_{CSi} = 1 - \delta_j + r_{CSi}, \text{ con } \delta_i \text{ como el indicador de no censura.}$$

Ejemplo 4.3.4. Los datos están asociados con el tiempo (en meses) hasta que el desgaste de las pastillas del sistema de frenado de un automóvil alcancen su punto límite y deban reemplazarse. La tabla 4.7 contiene los datos de tiempo con censura a derecha, pues el estudio se realizó durante 3 años (36 meses) contados a partir de la instalación de las pastillas y la puesta en marcha del vehículo por primera vez.

Los estimadores de máxima verosimilitud asociados con los parámetros son $\hat{\mu} = 3.12$ y $\hat{\sigma} = 0.51$. Se calculan los residuos de Cox-Snell $r_{CSi} = \ln S(t_i)$, con $S(t)$ como la función de sobrevida de la distribución log-normal. Una manera sencilla de calcular los residuales r_{CSi} es mediante la relación entre la distribución normal y la log-normal (3.5.3); esta es

$$S(t) = 1 - \Phi\left(\frac{\ln(t) - \mu}{\sigma}\right) = \Phi\left(\frac{-\ln(t) + \mu}{\sigma}\right).$$

Así, los residuos se calculan conforme a

$$r_{CSi} = -\ln[S(t_j)] = -\ln\left[1 - \Phi\left(\frac{\ln(t_j) - \mu}{\sigma}\right)\right].$$

Una vez calculados los residuos r_{CSi}, se obtiene el estimador de Kaplan-Meier de la función de sobrevida mediante $S_R(r_{CSi})$ y luego se computa $-\ln S_R(r_{CSi})$. Estos valores se consignan en la tabla 4.7.

Tabla 4.7. Tiempo a la falla en pastillas de frenos

$t_{(i)}$	$r_{(CSi)}$	$\hat{S}_R(r_{(CSi)})$	$-\ln \hat{S}_R(r_{(CSi)})$
	0.000	1.000	0.000
11	0.121	0.900	0.105
12	0.164	0.850	0.163
14	0.268	0.800	0.223
15	0.328	0.750	0.288
17	0.461	0.650	0.431
19	0.608	0.550	0.598
20	0.686	0.500	0.693
22	0.849	0.450	0.799
23	0.933	0.400	0.916
27	1.284	0.350	1.050
28	1.374	0.300	1.204
29	1.464	0.250	1.386
31	1.647	0.200	1.609
36	2.107	0.150	1.897
36^+	2.107		
42^+	2.658		
55^+	3.818		

Generado por el autor.

La gráfica de los residuos $r_{(CSi)}$ frente a $-\ln \hat{S}_R(r_{(CSi)})$, $i = 1, \ldots, 20$ muestra que la nube de puntos se ajusta bastante bien a una línea recta cuyo intersecto es cercano a cero y cuya pendiente es próxima a la unidad (inclinación de 45°). Por tanto, el modelo log-normal sigue o se ajusta adecuadamente a estos datos de tiempo hasta el desgaste en el tipo de pastillas de frenos consideradas en el estudio.

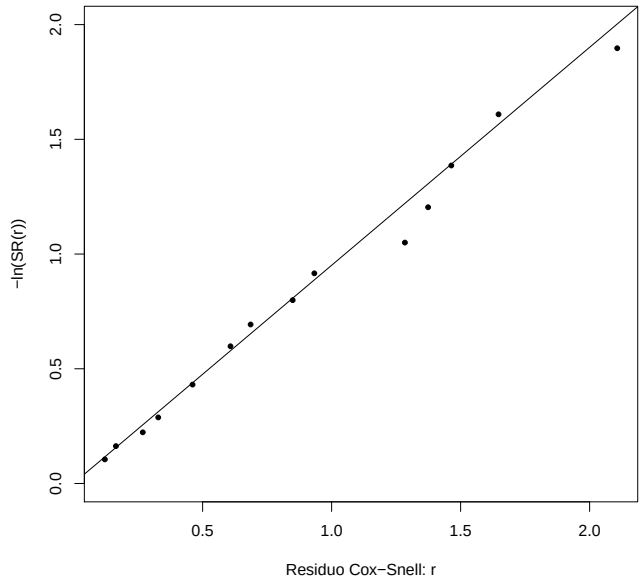

Figura 4.13. Gráfica de residuos de Cox-Snell para un modelo log-normal.

4.3.1. Prueba de bondad de ajuste

Además de las herramientas gráficas para la selección de modelos, se dispone de algunas estadísticas con las cuales se puede decidir sobre el ajuste de un modelo a los datos. Si se asume una distribución particular para los datos, esta queda definida por los parámetros, de manera que se debería desarrollar una verificación de hipótesis sobre un valor o conjunto de valores de los parámetros. Un segundo procedimiento puede consistir en asumir una familia de distribuciones, a la cual pertenece la distribución de interés para los datos. Por ejemplo, la distribución exponencial pertenece a la familia Weibull, y la Weibull, a su vez, pertenece a la familia gama generalizada.

Para el primer caso, en el cual se considera un modelo particular, $f(t, \theta)$, donde θ es un parámetro o un vector de p-parámetros, la distribución f es definida por θ, por lo que se pueden desarrollar verificaciones sobre algunos o todos los componentes de θ. Sea $\theta = (\theta_1, \theta_2)$ una partición del vector de p-parámetros (si θ es un escalar no hay lugar a la partición). Se quiere verificar la hipótesis nula

$$H_0 : \theta_2 = \theta_{20}. \tag{4.3.21}$$

Sean $\hat{\theta}$ el EMV de θ y $\hat{\theta}_1(\theta_0)$ el EMV de θ_1, dado que $\theta_2 = \theta_{20}$. La estadística de verificación de la hipótesis (4.3.21) es la estadística de razón de verosimilitud (RV):

$$RV = 2[l(\hat{\theta}) - l(\hat{\theta}_1(\theta_0), \theta_0)], \tag{4.3.22}$$

la cual tiene, de manera asintótica, una distribución ji-cuadrado con un número de grados de libertad igual al número de parámetros contenidos en θ_2.

Una alternativa para verificar la hipótesis (4.3.21) es la estadística de Wald:

$$W = (\hat{\theta}_2 - \theta_0)'\widehat{\Sigma}^{-1}(\hat{\theta}_2 - \theta_0), \tag{4.3.23}$$

donde $\widehat{\Sigma}$ es la matriz de covarianzas estimada para los componentes de θ_2. La estadística de Wald, W, tiene una distribución límite ji-cuadrado con un número de grados de libertad igual a la dimensión de θ_2.

En particular, si se quiere verificar la hipótesis de que el vector de parámetros es igual a un vector de valores conocidos, θ_0, la hipótesis nula tiene la forma

$$H_0 : \theta = \theta_0. \tag{4.3.24}$$

Mediante alguna de las siguientes estadísticas puede emplearse para verificar (4.3.24).

Razón de verosimilitud:

$$RV = 2[l(\hat{\theta}) - l(\theta_0)]. \tag{4.3.25}$$

Estadística de Wald:

$$W = (\hat{\theta} - \theta_0)' \left(\mathcal{IO}(\hat{\theta})\right) (\hat{\theta} - \theta_0), \tag{4.3.26}$$

donde

$$\mathcal{IO}(\hat{\theta}) = \left(-\frac{\partial^2 l(\hat{\theta})}{\partial \hat{\theta}_j \partial \hat{\theta}_k} \right)$$

es la matriz de información observada.

Estadística de puntuación:

$$\left(-\frac{\partial l(\theta_0)}{\partial \theta} \right)' \left(-\frac{\partial^2 l(\theta_0)}{\partial \hat{\theta}_j \partial \hat{\theta}_k} \right)^{-1} \left(-\frac{\partial l(\theta_0)}{\partial \theta} \right). \tag{4.3.27}$$

Bajo la hipótesis nula, $H_0 : \theta = \theta_0$, y como $\hat{\theta}$ tiene distribución asintótica normal p-variante, cada una de estas tres estadísticas tiene distribución asintótica ji-cuadrado con p grados de libertad (la dimensión de θ).

Se debe tener precaución con la interpretación del rechazo de H_0, pues significa que la distribución generada por los parámetros θ_0 no es apropiada para los datos observados; esto no significa que sea inapropiada la familia de distribuciones a la que pertenece la distribución de interés. Es posible que una distribución de la misma familia pero con diferente θ_0 sea apropiada.

El segundo tipo de hipótesis implica identificar una familia de modelos tal que la distribución de interés sea un caso particular o un miembro de esa familia (modelos encajados). El procedimiento de verificación es como sigue:

1. Definir la familia general de modelos y obtener el logaritmo de la verosimilitud, este es, $l_G = \ln L(\hat{\theta}_G)$.

2. Especificar el modelo de interés o el propuesto para los datos, es decir, establecer la hipótesis nula H_0: "El modelo propuesto es adecuado para los datos". De igual manera, calcular el logaritmo de su verosimilitud, es decir, $l_P = \ln L(\hat{\theta}_P)$.

3. Con estas dos expresiones se calcula la razón de verosimilitud,

$$-2 \ln \left(\frac{L(\hat{\theta}_P)}{L(\hat{\theta}_G)} \right) = 2[l_G(\hat{\theta}_G) - l_P(\hat{\theta}_P)], \tag{4.3.28}$$

   estadística que, bajo H_0, tiene una distribución límite ji-cuadrado con un número de grados de libertad igual a la diferencia entre el número de parámetros de θ_G y θ_P.

Capítulo

cinco

Modelos de regresión paramétricos en A. S.

$h_X(t) = h_0(t)g_1(x)$

5.1. Introducción

Hasta ahora, en el tratamiento que se ha dado al modelamiento de la variable *tiempo hasta la ocurrencia del evento*, ya sea mediante su función de sobrevida o de confiabilidad y su función de riesgo, no se han considerado algunos factores y variables que tienen que ver con la unidad o su entorno. No obstante, la ocurrencia del evento, ya sea de manera temprana o tardía, puede atribuirse al efecto de algunas covariables ligadas a la unidad o su entorno.

Por ejemplo, en el estudio del tiempo para que un niño aprenda a leer o escribir, se pueden buscar factores asociados tales como las características del niño (antropométricas, psicológicas y demográficas), el entorno familiar del niño (aspectos socioeconómicos, demográficos y culturales) y las características de la institución educativa (profesores, planta física, prácticas de aula). Para el caso de tiempo hasta la remisión de un paciente con alguna dolencia, las variables están contenidas en la ficha o historia clínica del paciente, junto con algunos aspectos del procedimiento médico y clínico. En un motor que funciona con combustible, el tiempo hasta la falla puede atribuirse al sistema de lubricación, de refrigeración, de alimentación o al eléctrico, así como también a algunos aspectos ligados al operario y al ambiente de operación.

En las situaciones descritas, además de la variable aleatoria *tiempo para el evento*, se dispone de un conjunto de covariables, las cuales el investigador postula o hipotetiza como influyentes en la variable *tiempo al evento*. Así, se considera un vector de covariables $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$. Para valores específicos de estas variables, medidos sobre alguna unidad particular, se escribe $\mathbf{X} = x = (x_1, x_2, \ldots, x_p)'$. Estas variables pueden ser de tipo cuantitativo o categórico. Por ejemplo, en el caso de remisión de un paciente, las variables *estatura* y *peso* se consideran como cuantitativas, mientras que variables como *sexo*, *tipo de ocupación* o *profesión del paciente* se consideran como categóricas; de manera que el vector $\mathbf{X}$ puede contener ambos tipos de variables.

En este capítulo se estudia la posible asociación entre la variable *tiempo para el evento* y un conjunto de variables explicativas (covariables). Estas variables, notadas por $\mathbf{T}$ y $\mathbf{X}$, se escriben dentro de un modelo estadístico, el cual relaciona una función de la variable $\mathbf{T}$ con una función de las variables contenidas en $\mathbf{X}$. El modelo es una expresión matemática que contiene las formas funcionales de las dos variables, relacionadas o "conectadas" mediante una igualdad.

Los modelos en los cuales se asume una distribución probabilística específica para la variable aleatoria *tiempo hasta el evento*, $\mathbf{T}$, se conocen como

modelos paramétricos. La distribución central en el análisis de datos de tiempo hasta un evento es la distribución Weibull, así como en los modelos lineales la distribución normal tiene un papel central.

Dentro de los modelos paramétricos se presentan dos clases, los modelos de *riesgos proporcionales* y los modelos de *tiempo acelerado*. Este capítulo está orientado por Lee y Wang [44], Colosimo y Giolo [11], Klein y Moeschberger [39] y Huan.

5.2. Modelos de riesgos proporcionales

La variable *tiempo hasta el evento*, $\mathbf{T}$, se modela como dependiente del vector de covariables o variables explicativas, $\mathbf{X}$; se escribe $\mathbf{T_X}$ para representar la posible dependencia estadística entre estas dos variables. En consecuencia, la función de riesgo también dependerá de estas covariables. Se escribe $h_{\mathbf{x}}(t)$ para indicar la función de riesgo de T cuando $\mathbf{x}$ es un vector de covariables observadas.

Así, se trata de dar cuenta del efecto de estas covariables sobre la función de riesgo y de la forma en que ese efecto debe considerarse e incluirse en el modelo. Se asume que las covariables tienen un efecto multiplicativo sobre una función denominada *riesgo base*.

Un *modelo de riesgos proporcionales* para $\mathbf{T_X}$ significa

$$h_{\mathbf{X}}(t) = h_0(t)g_1(\mathbf{x}), \tag{5.2.1}$$

donde $g(\cdot)$ es una función positiva en $\mathbf{X}$ y $h_0(t)$ es la función de riesgo base, la cual representa la función de riesgo en ausencia de covariables, o en caso de tener covariables, todas tendrían el mismo efecto sobre el riesgo, es decir, $g_1(x) = 1$.

Así, en un *modelo de riesgos proporcionales* las covariables tienen un efecto multiplicativo sobre el riesgo a través del producto con $h_0(t)$. Esta función de riesgo base, $h_0(t)$, no depende de $\mathbf{X}$, mientras que g_1 depende de $\mathbf{X}$. El calificativo de *proporcional* se puede explicar de esta manera: si dos unidades tienen tiempos hasta el evento dependientes de los respectivos valores en las covariables $\mathbf{x}_1$ y $\mathbf{x}_2$, entonces

$$\frac{h_{\mathbf{x}_1}}{h_{\mathbf{x}_2}} = \frac{h_0(t)g_1(\mathbf{x}_1)}{h_0(t)g_1(\mathbf{x}_2)} = \frac{g_1(\mathbf{x}_1)}{g_1(\mathbf{x}_2)}.$$

Este resultado muestra que la razón de riesgos para dos unidades no depende del tiempo t sino de los valores que toman las covariables en las respectivas unidades.

La función $g_1(\cdot)$ debe ser una función de valores no negativos, es decir, $g_1(x) \geq 0$ y $g_1(0) = 1$. Generalmente,

$$g_1(\mathbf{x}) = e^{\{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\}} = e^{(x'\beta)}, \tag{5.2.2}$$

donde $\beta_j, j = 1, 2, \ldots, p$, es el coeficiente de la variable X_j, que se interpreta como el efecto parcial de esta variable sobre la tasa de riesgo del evento. El modelo de riesgos proporcionales se puede escribir como

$$h_{\mathbf{x}}(t) = h(t \mid \mathbf{x}) = h_0(t)e^{(\mathbf{x}'\beta)}. \tag{5.2.3}$$

Se introduce la notación $h(t \mid \mathbf{x})$ para hacer más explícita la consideración de las covariables como posibles variables explicativas de la tasa de riesgo. Aunque matemáticamente $h(t \mid \mathbf{x}) = h_0(t)$ cuando $\mathbf{x} = 0$, se debe tener cuidado para el caso en el cual $\mathbf{x}$ tenga algunas variables categóricas o estrictamente nominales, de otra manera, $\mathbf{x} = 0$ es una forma de considerar el riesgo sin covariables. La estimación del modelo expresado en (5.2.3) implica la estimación de los p parámetros contenidos en el vector β, los cuales se obtienen desde los tiempos observados y los valores de las covariables.

Ahora, para la función de sobrevida o de confiabilidad, se escribe $S(t|\mathbf{x})$ para indicar la función de sobrevida para la unidad definida por el perfil de valores contenidos en el vector x. Se sigue entonces, desde la igualdad (1.5.4), que

$$\begin{aligned} S(t|\mathbf{x}) &= \exp\left(-\int_0^t h(u \mid \mathbf{x})du\right) \\ &= \exp\left(-\int_0^t h_0(u)g_1(\mathbf{x})du\right) \\ &= \exp\left(-g_1(\mathbf{x})\int_0^t h_0(u)du\right) \\ &= \left[\exp\left(-\int_0^t h_0(u)du\right)\right]^{g_1(\mathbf{x})} \\ &= [S_0(t)]^{g_1(\mathbf{x})}, \end{aligned} \tag{5.2.4}$$

donde $S_0(t) = \exp\left(-\int_0^t h_0(u)du\right)$ es la función de sobrevida base, la cual se tiene cuando $\mathbf{x} = 0$ o en ausencia de covariables. Una consecuencia del supuesto del modelo de riesgo proporcional sobre $S(t|\mathbf{x})$ es, desde (5.2.4), que especifica una función $g_1(\cdot)$ y una función de sobrevida $S_0(t)$ tal que

$$S(t|\mathbf{x}) = [S_0(t)]^{g_1(\mathbf{x})}.$$

Ejemplo 5.2.1. Suponga que se tiene como única covariable $\mathbf{x}(p = 1)$, que toma los valores 1 o 0 según la unidad pertenezca o no a un grupo de interés (por ejemplo, un tratamiento o grupo control).

La función $S_0(t)$ se refiere a la función de sobrevida para las unidades que no pertenecen al grupo de interés. Así,

$$g_1(\mathbf{x}) = e^{\mathbf{x}'\beta} = e^{\beta} = \gamma.$$

Es decir, $g_1(1) = \gamma > 0$. Análogamente, si x es el valor de la covariable para una unidad fuera del grupo de interés, entonces $g_1(\mathbf{x}) = e^{\mathbf{x}'\beta} = e^0 = 1$, por lo que la función de sobrevida para el grupo de interés, $\mathbf{x} = 1$, es

$$S_1(t) = S_0(t)^{\gamma}, \tag{5.2.5}$$

la cual define un conjunto de distribuciones llamada *familia de Lehmann*.

Cuando el vector X se ha registrado sobre un número de n unidades, se representa por $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ el valor de cada una de las p variables sobre la unidad $i = 1, 2, \ldots, n$. Para el modelo, se debe considerar la variable *tiempo al evento* para la unidad i, $\mathbf{T}_i$, junto con su función de sobrevida S_i, función de probabilidad f_i y de riesgo h_i dadas por

$$S_i(t) = S_{\mathbf{x}_i},\ f_i = f_{\mathbf{x}_i}(t),\ h_i(t) = h_{\mathbf{x}_i}(t).$$

El riesgo para la i-ésima unidad es

$$\frac{h_i(t)}{h_0(t)} = e^{\mathbf{x}_i}\beta.$$

Se concluye, desde este cociente o razón, que la función de riesgo para una unidad $i = 1, 2, \ldots, n$, en particular, es un múltiplo diferente del riesgo base. El múltiplo es diferente porque cada unidad toma valores específicos en el vector de covariables X. Mediante el modelamiento, se estima este múltiplo a partir de la estimación de los parámetros contenidos en β, el vector de los efectos de las covariables.

Modelo de riesgo proporcional para datos Weibull

La función de sobrevida para una variable aleatoria $Weibull(\theta, \alpha)$, de acuerdo con la ecuación (3.4.3), es

$$S(t) = e^{-(\theta t)^{\alpha}},\ t \geq 0.$$

Para la construcción de un modelo de regresión con este modelo probabilístico se debe expresar el parámetro de escala, θ, como una combinación

lineal de las covariables. Es decir, el parámetro de escala se pone ahora como un funcional de las covariables, este es, $\theta(\mathbf{x})$. Así, la función de sobrevida es ahora modelada como

$$S(t|\mathbf{x}) = e^{-[\theta(\mathbf{x})t]^{\alpha}}.$$

Esta estructura de la función de sobrevida con las covariables no es arbitraria, pues debe ser de esta forma para tener una estructura de modelo de riesgo proporcional. A continuación se muestra que esta estructura de la función de sobrevida, del tiempo hasta el evento y de las covariables es adecuada para un modelo Weibull. Se quiere usar la función $g_1(\mathbf{x})$ de la forma

$$g_1(\mathbf{x}) = e^{\mathbf{x}'\beta}.$$

Suponga que se escoge $\theta(\mathbf{x}) = g_1(\mathbf{x})^{\frac{1}{\alpha}}$. La función de riesgo Weibull, de acuerdo con (3.4.4), es

$$h(t|\mathbf{x}) = \theta(\mathbf{x})^{\alpha}\alpha(t)^{\alpha-1} = [\alpha t^{\alpha-1}]\theta(\mathbf{x})^{\alpha} = h_0(t)g_1(\mathbf{x}).$$

De la definición de modelos de riesgos proporcionales dada por la expresión (5.2.1), se tiene que $h_0(t) = \alpha t^{\alpha-1}$. Nuevamente por (3.4.4), esta es la función de riesgo asociada con una distribución $Weibull(1, \alpha)$ de la variable tiempo hasta el evento, la cual no depende de las covariables, y $g_1(\mathbf{x}) = \theta(\mathbf{x})^{\alpha}$.

En conclusión, si se ajusta un modelo Weibull a las covariables contenidas en $\mathbf{X}$, entonces

$$\mathbf{T_X} \sim Weibull(\theta(\mathbf{x}), \alpha),$$

la cual tiene asociada una función de riesgos proporcionales, cuya función de riesgo base es $h_0(t) = \alpha t^{\alpha-1}$.

La función de riesgos proporcionales para la unidad i, identificada por sus valores en las covariables $\mathbf{x_i}$, con $g_1(\mathbf{x}) = e^{\mathbf{x}'\beta}$, es

$$h(t|\mathbf{x_i}) = h_0(t)e^{\mathbf{x_i}'\beta}.$$

La estimación de los parámetros α y β se hace mediante el método de máxima verosimilitud.

5.3. Modelos de tiempo acelerado

En estos modelos se asume que las covariables contenidas en el vector X actúan directamente sobre el tiempo al evento y, por lo tanto, aceleran o retrasan el tiempo para el evento de interés. Para algunos eventos se busca

que el tiempo para su ocurrencia sea largo (retraso), mientras que en otros se intenta precipitar (acelerar) su ocurrencia; en este sentido, se debe explorar y confirmar cuáles de las covariables evidencian este propósito. Por ejemplo, en los eventos como el de una enfermedad, aprender a leer o adquirir un empleo, se deben buscar los factores o covariables asociables con la anticipación o precipitación del evento. Las demás covariables pueden o no tener influencia sobre el tiempo para el evento o, por el contrario, provocar su retraso. En otra dirección, en eventos como la muerte, la falla o la recurrencia de una enfermedad, se buscan las covariables que retarden la ocurrencia de estos. A estas variables o factores se les denomina *factores de riesgo* o *factores protectores*, según se encuentre su asociación con la demora o la aceleración en la ocurrencia de un evento de interés.

Los modelos de tiempo acelerados para el evento relacionan la distribución de la variable *tiempo* con las variables explicativas, también denominadas variables de "estrés", covariables o regresores. El efecto de las covariables se puede evidenciar en la función de sobrevida, la distribución acumulada o las funciones de densidad de probabilidad. Sin embargo, el sentido de los modelos de tiempo acelerados se evidencia mejor si se formulan en términos de la función de tasa de riesgo instantáneo.

En términos de la variable *tiempo hasta el evento*, la precipitud o tardanza del evento de interés se modela multiplicando la variable *tiempo hasta el evento* por una función de las covariables, $g_2(\mathbf{x})$, es decir,

$$\mathbf{T_X} g_2(\mathbf{x}) = \mathbf{T}_0, \text{ o equivalentemente } \mathbf{T_X} = \frac{\mathbf{T}_0}{g_2(\mathbf{x})}. \tag{5.3.1}$$

La pregunta que se debe responder es: ¿cómo afectan las covariables la función de sobrevida de $\mathbf{T_X}$? Una respuesta se tiene con el cálculo de la función de sobrevida

$$S(t|\mathbf{x}) = P(\mathbf{T_X} > t) = P\left(\frac{\mathbf{T}_0}{g_2(\mathbf{x})} > t\right) = P(\mathbf{T}_0 > t g_2(\mathbf{x})) = S_0(t g_2(\mathbf{x})). \tag{5.3.2}$$

Esta igualdad es la que caracteriza un modelo de tiempo acelerado.

Un *modelo de tiempo para un evento acelerado* se expresa mediante el modelo

$$S(t|\mathbf{x}) = S_0[t g_2(\mathbf{x})], \tag{5.3.3}$$

donde $g_2(\mathbf{x})$ es una función de las covariables positiva. A $S_0(t)$ se le denomina *función de sobrevida base*, y representa la función de sobrevida de una unidad para la cual $g_2(\mathbf{x}) = 1$.

La selección usual de g_2 es $g_2(\mathbf{x}) = e^{-\mathbf{x}'\beta}$, selección que es conveniente, pues $g_2(\mathbf{x}) \geq 0$ para todo x y $g_2(0) = 1$. De aquí se puede derivar que

$$\mathbf{T_X} = \mathbf{T}_0 e^{\mathbf{x}'\beta}. \tag{5.3.4}$$

Si $\mathbf{x}'\beta < 0$, entonces a las covariables se les puede atribuir que estarían asociadas con la contracción desde $\mathbf{T}_0$ a $\mathbf{T_X}$ (precipitud de la ocurrencia del evento), mientras que si $\mathbf{x}'\beta > 0$, entonces a las covariables se les atribuye que el tiempo al evento se prolongue desde $\mathbf{T}_0$ a $\mathbf{T_X}$ (retraso de la ocurrencia del evento).

Tomando el logaritmo natural en los dos miembros de la igualdad (5.3.1) y considerando que $g_2(\mathbf{x}) = e^{-\mathbf{x}'\beta}$, se tiene

$$\begin{aligned} \ln \mathbf{T_X} &= \ln \mathbf{T}_0 - \ln g_2(\mathbf{x}) \\ &= \ln \mathbf{T}_0 + \mathbf{x}'\beta \\ &= \mu_0 + \mathbf{x}'\beta + \sigma Z, \end{aligned} \tag{5.3.5}$$

donde $\ln \mathbf{T}_0 = \mu_0 + \sigma Z$, con Z como una variable aleatoria de media igual a 0 y varianza igual a 1. Esto implica que se dispone de un *modelo log-lineal* de la forma

$$\ln \mathbf{T_X} = \mu(\mathbf{x}) + \sigma Z, \tag{5.3.6}$$

donde $\mu(\mathbf{x}) = \mu_0 + \mathbf{x}'\beta$. A partir de la identidad (1.5.2) y por la definición de modelo acelerado contenida en (5.3.3), derivando la función de sobrevida $S(t|\mathbf{x})$ respecto de t, conforme se indica en (1.5.3), se encuentra que

$$h(t|\mathbf{x}) = h_0[t g_2(\mathbf{x})] g_2(\mathbf{x}). \tag{5.3.7}$$

Modelo de tiempo acelerado para datos Weibull

Para modelos de tiempo acelerados, el logaritmo del tiempo para el evento es una función lineal tanto de las covariables como de los parámetros. Para la distribución Weibull, la función de sobrevida, de acuerdo con (3.4.3), es

$$S(t) = 1 - F(t) = e^{-(\theta t)^{\alpha}}.$$

De manera que, si se incluyen las covariables en lugar del parámetro de escala θ, es decir, se reemplaza por $\theta(\mathbf{x})$, entonces, si $\mathbf{T_X}$ tiene una distribución $Weibull(\theta(\mathbf{x}), \alpha)$, la función de sobrevida de $\mathbf{T_X}$ es

$$S(t|\mathbf{x}) = \exp\left[-(\theta(\mathbf{x})t)^{\alpha}\right].$$

A partir de una adecuada escogencia de la función de los parámetros, $\theta(\mathbf{x})$, se debe obtener el modelo log-lineal como en (5.3.6). Para este propósito, considere

$$\theta(\mathbf{x}) = g_2(\mathbf{x}) = e^{\mathbf{x}'\beta}.$$

Se sigue, a partir de la derivación de la función de sobrevida *de valor extremo* [64, p. 27], que si T tiene distribución $Weibull(\theta, \alpha)$, entonces

$$\ln \mathbf{T} \sim VE(u, b), \text{ con } u = \ln\theta \text{ y } b = \frac{1}{\alpha}, \tag{5.3.8}$$

y mediante una "estandarización" de $\ln \mathbf{T}$, se tiene que

$$\frac{\ln \mathbf{T} - u}{b} \sim VE(0, 1). \tag{5.3.9}$$

Para el modelo de tiempo acelerado, es decir, con la variable *tiempo al evento* "influida o afectada" por las covariables, $\mathbf{T_X}$, se tiene que

$$\mathbf{Z} = \frac{\ln \mathbf{T_X} - u(\mathbf{x})}{b} \sim VE(0, 1). \tag{5.3.10}$$

Para $u(\mathbf{x}) = \ln\theta(\mathbf{x}) = \mathbf{x}'\beta$ y $b = \frac{1}{\alpha}$, las expresiones (5.3.9) y (5.3.10) representan el mismo modelo cuando $\mathbf{Z}$ tiene una distribución de valor extremo con parámetros 0 y 1 y un parámetro de escala $\sigma = b = \frac{1}{\alpha}$.

En consecuencia, cuando el modelo log-lineal contenido en (5.3.9) es aplicado a un conjunto de datos Weibull, se tiene que

$$\ln \mathbf{T_X} = \mu_0 + \mathbf{x}'\beta + \sigma Z, \text{ para } \mathbf{Z} \sim VE(0, 1). \tag{5.3.11}$$

Note el paralelo de $\mathbf{Z}$ con los errores de un modelo lineal clásico, pues en ambos se asumen un distribución normal con media 0 y varianza σ^2.

Los *modelos de riesgos proporcionales* y los modelos *acelerados sobre el tiempo para el evento* coinciden si y solo si los tiempos para el evento tienen distribución Weibull.

La argumentación de esta proposición consiste en demostrar que las definiciones de los modelos coinciden para la distribución Weibull de la variable *tiempo hasta el evento*. Es decir, que (5.2.1) y (5.3.3) coinciden, lo cual es equivalente a que se satisfaga la igualdad $S(t|\mathbf{x}) = [S_0(t)]^{g_1(\mathbf{x})}$ si y solo si existen $\theta > 0$ y $\alpha > 0$, tal que $S_0(t) = e^{-(\theta t)^\alpha}$. La demostración completa se puede consultar en Smith [64]. A continuación se presentan algunos modelos de regresión paramétricos particulares.

5.3.1. Modelo lineal para datos de tiempo hasta un evento

Un modelo de regresión es una relación que se puede establecer entre variables. Una de las variables se considera como respuesta o dependiente, por decisión o interés del investigador, y las demás se asumen como variables explicativas de la variable respuesta. La variable respuesta se escribe como una función de las variables explicativas; la forma funcional depende de los datos y, más específicamente, del marco conceptual en el que se inscriben los datos. Una representación gráfica de la variable respuesta frente a cada una de las variables explicativas (covariables o regresores) coadyuva a explorar y considerar la relación matemática o la forma funcional (una línea recta, una parábola, un polinomio) entre estas.

Para el caso del análisis de la variable *tiempo hasta la ocurrencia de un evento*, $\mathbf{T}$, y las covariables $\mathbf{X} = (1, X_1, X_2, \ldots, X_p)'$, una representación funcional de la relación entre la variable respuesta y las covariables es

$$\mathbf{T} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon = \mathbf{X}'\beta + \epsilon, \qquad (5.3.12)$$

donde $\mathbf{T}$ es la variable respuesta, el vector X contiene las covariables X_i, $i = 1, 2, \ldots, p$ y $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ son los parámetros que deben ser estimados desde los datos y representan el efecto de cada covariable sobre $\mathbf{T}$. Además, ϵ es un error aleatorio con distribución normal. Al componente $\mathbf{X}'\beta$ se le denomina *determinístico* y a ϵ, el componente *aleatorio*.

La variable respuesta *tiempo hasta la ocurrencia de un evento* generalmente presenta censura en algunos de sus valores. Además, la distribución de la respuesta tiende a ser asimétrica en dirección de los mayores tiempos. Estas características hacen inapropiado el uso de la distribución normal para el componente aleatorio ϵ. En la metodología del análisis de regresión existen varias estrategias para tratar este tipo de datos; se destacan las dos siguientes:

1. Transformar la respuesta, $\mathbf{T}$, para alcanzar una distribución normal. La transformación de Box-Cox es una opción de transformación.

2. Utilizar un componente determinístico no lineal en los parámetros para la asociación entre las covariables y considerar una distribución asimétrica para el componente aleatorio ϵ.

Estas dos opciones pueden ser equivalentes. Por ejemplo, emplear un modelo lineal para la transformación logarítmica de la respuesta equivale a emplear un componente determinístico $\exp(\mathbf{X}'\beta)$ y la distribución lognormal para el error.

A continuación, se presentan algunos distribuciones asimétricas para el error.

Modelo de regresión exponencial

Se emplea la distribución exponencial para el error y el componente determinístico se considera de la forma $\exp(\mathbf{X}'\beta)$. Una combinación del componente determinístico con una distribución exponencial para el error, cuya media es la unidad (con distribución $f(\epsilon) = e^{-\epsilon}$), genera el siguiente modelo:

$$\mathbf{T} = \exp\{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\}\epsilon = \exp\{\mathbf{x}'\beta\}\epsilon. \qquad (5.3.13)$$

Este es el modelo de regresión exponencial. Se observa que la relación entre $\mathbf{T}$ y el componente determinístico es no lineal, además, el componente aleatorio tiene una distribución asimétrica.

El modelo (5.3.28) es linealizable, pues si se aplican logaritmos en los dos lados de la igualdad se obtiene

$$\ln \mathbf{T} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \ + \upsilon, \qquad (5.3.14)$$

con $\upsilon = \ln \epsilon$. El modelo (5.3.14) es un modelo lineal con una distribución de los errores, υ, que no tiene una distribución normal, pues su función de densidad de probabilidad es $f(\upsilon) = \exp[\upsilon - \exp(\upsilon)]$, la cual es de la familia de distribuciones de valor extremo (distribución Gumbel). La utilidad de esta distribución es la aplicación del logaritmo (para "atraer" valores lejanos o extremos) sobre ciertos tiempos para el evento, y en consecuencia, la distribución de los datos transformados se adecua a los datos.

Se nota que en la transformación logarítmica los $\mathbf{X}$ actúan aditivamente sobre $\ln \mathbf{T}$, ecuación (5.3.14), mientras que en la ecuación (5.3.21) actúan multiplicativamente.

La función de sobrevida condicionada a las covariables $\mathbf{x}$ se calcula mediante

$$S(t|\mathbf{x}) = \exp\left[-\exp\{\mathbf{x}'\beta\}t\right]. \qquad (5.3.15)$$

La función de densidad de probabilidad es

$$f(t|\mathbf{x}) = \exp(-\mathbf{x}'\beta)\exp[-\exp(-\mathbf{x}'\beta)t], \qquad (5.3.16)$$

la función de riesgo, calculada mediante (1.5.2), viene dada por

$$h(t|\mathbf{x}) = \exp(-\mathbf{x}'\beta). \qquad (5.3.17)$$

Así, por la expresión (5.3.17), el modelo de regresión exponencial asume una relación lineal entre las covariables y el logaritmo del riesgo.

Suponga que las unidades i y j se caracterizan por los valores que tomen en el vector de covariables, $\mathbf{x}_i$ y $\mathbf{x}_j$, respectivamente; la razón de riesgo de estas unidades es

$$\frac{h(t|\mathbf{x}_i)}{h(t|\mathbf{x_j})} = \exp\left[-\sum_{k=1}^{p}\beta_k(x_{ki}-x_{kj})\right].$$

Esta razón de riesgos solo depende de las diferencias entre las respectivas covariables observadas sobre las unidades y los coeficientes de regresión. No depende del tiempo para el evento, y en consecuencia, se puede considerar al modelo de regresión exponencial como un modelo de riesgos proporcionales, en el cual la razón de riesgo de cualquier par de unidades es independiente del tiempo y constante para cada par. El modelo de regresión exponencial es, entonces, un caso especial de los modelos de riesgos proporcionales.

La función de verosimilitud es

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n}\left(f(t_i|\mathbf{x_i})\right)^{\delta_i}\left(S(t_i|\mathbf{x_i})\right)^{1-\delta_i} \\ &= \prod_{i=1}^{n}\left(\exp(-\mathbf{x_i'}\beta)\exp[-\exp(-\mathbf{x_i'}\beta)t_i]\right)^{\delta_i}\left(\exp\left[-\exp\{\mathbf{x'}\beta t_i\}\right]\right)^{1-\delta_i} \\ &= \prod_{i=1}^{n}\left(\exp(-\mathbf{x_i'}\beta)\right)^{\delta_i}\exp[-\exp(-\mathbf{x_i'}\beta)t_i]. \end{aligned} \tag{5.3.18}$$

Tomando logaritmos en los dos miembros de (5.3.18) se obtiene

$$l(\beta) = -\sum_{i=1}^{n}\left[\delta_i(\mathbf{x_i'}\beta)) + \exp(-\mathbf{x_i'}\beta)t_i\right]. \tag{5.3.19}$$

Los estimadores de máxima verosimilitud para el vector de parámetros de la regresión, β, se obtienen al solucionar el sistema de ecuaciones

$$\begin{aligned} \frac{\partial l(\beta)}{\partial \beta} &= 0 \\ \sum_{i=1}^{n}\left[\delta_i(\mathbf{x_i})) - t_i\mathbf{x_i}\exp(-\mathbf{x_i'}\beta)\right] &= 0. \end{aligned} \tag{5.3.20}$$

Este sistema de ecuaciones se puede resolver mediante el método de Newton-Raphson: la inferencia para los EMV de β se puede hacer mediante las estadísticas de Wald, la razón de verosimilitud o la estadística de puntuación, como se muestra en la sección 4.2.

Modelo de regresión Weibull

De manera similar a la distribución exponencial, se considera la expresión $\exp(\mathbf{x}'\beta)$, que es equivalente al modelo

$$\ln \mathbf{T} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \ + \upsilon, \tag{5.3.21}$$

donde υ tiene la distribución definida para el modelo anterior. En el modelo de Weibull, $\mathbf{T}$ tiene una distribución Weibull, entonces, las funciónes de densidad de probabilidad, de sobrevida y riesgo están dadas, respectivamente, por:

$$\begin{aligned} f(t|\mathbf{x}) &= \alpha t^{\alpha-1} \exp(-\alpha \mathbf{x}'\beta) \exp[-\exp(-\alpha \mathbf{x}'\beta) t^{\alpha}] \\ S(t|\mathbf{x}) &= \exp\left[-\exp\{-\alpha \mathbf{x}'\beta\} t^{\alpha}\right] \\ h(t|\mathbf{x}) &= \alpha \exp(-\alpha \mathbf{x}'\beta) t^{\alpha-1}. \end{aligned} \tag{5.3.22}$$

La razón de riesgo entre cualquier par de unidades, i y j, de la población objeto de estudio, las cuales se caracterizan por los valores que tomen en el vector de covariables, $\mathbf{x}_i$ y $\mathbf{x}_j$, respectivamente, se define, de acuerdo con (5.3.22), como

$$\frac{h(t|\mathbf{x}_i)}{h(t|\mathbf{x_j})} = \exp\left[-\alpha \sum_{k=1}^{p} \beta_k (x_{ki} - x_{kj})\right].$$

Esta razón es independiente del tiempo. Entonces como en el modelo de regresión exponencial, el modelo de regresión Weibull es también un modelo de riesgos proporcionales.

La estimación de los parámetros del predictor lineal $\mathbf{x}'\beta$ se obtienen mediante el método de máxima verosimilitud. La función de verosimilitud, de acuerdo con (5.3.22), es

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} \left(f(t_i|\mathbf{x_i})\right)^{\delta_i} \left(S(t_i|\mathbf{x_i})\right)^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left(\alpha t_i^{\alpha-1} \exp(-\alpha \mathbf{x}'\beta)\right)^{\delta_i} \exp[-\exp(-\alpha \mathbf{x}'\beta) t_i^{\alpha}]. \end{aligned} \tag{5.3.23}$$

Para simplificar los cálculos, se toman logaritmos a la función de verosimilitud, y el resultado es

$$\ln L(\beta) = l(\beta) = \sum_{i=1}^{n} \{\delta_i [\ln \alpha + (\alpha - 1) \ln t_i - \alpha \mathbf{x}'\beta] - \exp(-\alpha \mathbf{x}'\beta) t^{\alpha}\}. \tag{5.3.24}$$

Como se observa en el logaritmo de la función de verosimilitud (5.3.24), los parámetros a estimar son α y los contenidos en el vector β; no obstante, se escribe $l(\beta)$ para resaltar que el objetivo es estimar el modelo de regresión Weibull. Hay dos alternativas para la estimación. La primera sería estimar el parámetro α y emplear esta estimación en el logaritmo de la verosimilitud (5.3.24), y entonces estimar solamente los β. La segunda alternativa es, a partir del logaritmo de la verosimilitud (5.3.24), resolver el sistema de ecuaciones:

$$\frac{\partial l(\alpha, \beta)}{\partial \alpha} = 0 \;\; y \;\; \frac{\partial l(\alpha, \beta)}{\partial \beta} = 0,$$

solución que suministra los EMV $\hat{\alpha}$ y $\hat{\beta}$. La alternativa a seguir depende de la información y la herramienta computacional disponibles.

Modelo de regresión log-normal

Para este modelo se asume que el modelo de regresión para la variable tiempo hasta la ocurrencia de un evento, $\mathbf{T}$, es de la forma de (5.3.28), es decir,

$$\ln \mathbf{T} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p x_p \; + \upsilon$$

es el modelo de regresión log-normal. Así, $\mathbf{T}$ tiene distribución log-normal con función de densidad de probabilidad

$$f(t|\mathbf{x}) = \frac{\exp[-(\ln t - \mathbf{x}'\beta)^2/2\sigma^2]}{\sqrt{2\pi}\sigma t}. \tag{5.3.25}$$

La función de sobrevida es

$$S(t|\mathbf{x}) = 1 - \Phi\left(\frac{\ln t - \mathbf{x}'\beta)}{\sigma}\right). \tag{5.3.26}$$

Los coeficientes de regresión β y el parámetro σ deben estimarse desde los datos de tiempo y de las covariables; esto se hace mediante el método de máxima verosimilitud. Para datos de tiempo al evento con censura a derecha, el logaritmo de la verosimilitud es

$$l(\beta, \sigma) = \sum_{nc}\left[-\frac{(\ln t_i - \mathbf{x}'\beta)^2}{2\sigma^2} - \ln(\sqrt{2\pi}\sigma t_i)\right] + \sum_{c}\left\{\ln\left[1 - \Phi\left(\frac{\ln t_i - \mathbf{x}'\beta}{\sigma}\right)\right]\right\}, \tag{5.3.27}$$

donde los contadores de las sumas NC y C significan tiempo no censurado y censurado, respectivamente.

Los EMV de los parámetros contenidos en el vector β y σ se encuentran como soluciones de las ecuaciones

$$\frac{\partial l(\sigma, \beta)}{\partial \beta} = 0 \ \ y \ \ \frac{\partial l(\sigma, \beta)}{\partial \sigma} = 0.$$

Un procedimiento iterativo para resolver este sistema de ecuaciones es el de Newton-Raphson, como se indica en la expresión (4.2.8). Los procedimientos para la verificación de hipótesis sobre los β se pueden desarrollar mediante la estadística de Wald, la razón de verosimilitud o por medio de la estadística de puntuación, como se presenta en la sección 4.2.

Ejemplo 5.3.1. En la tabla 5.1 se consignan los tiempos de sobrevivencia para 33 pacientes con leucemia. Los tiempos se miden en semanas a partir del diagnóstico de la leucemia. Además se tienen dos covariables: el conteo de glóbulos blancos (CGB) en el momento del diagnóstico y una variable binaria (AG) que indica presencia (positiva con $AG = 1$) o ausencia (negativa con $AG = 0$) de una prueba sobre las características de las células de los glóbulos blancos. Los datos originales no tenían tiempos de supervivencia censurados, pero con fines ilustrativos, tres de los tiempos han sido reemplazados con tiempos censurados. En aspectos metodológicos se sigue a Colosimo y Giolo [11, pp. 129-140]) y a Lawless [43, pp. 276-277]. Con el propósito de mostrar la metodología de estructura, estimación y ajuste de un modelo de regresión paramétrico, se hace el análisis de regresión sobre los datos asociados a los 17 pacientes con diagnóstico positivo (Diag. = 1) de la tabla 5.1.

La variable CGB se considera continua (en unidades de 1000). Se pretende un análisis de regresión para estos datos con covariable CGB para los 17 pacientes que tuvieron una caracterización positiva de los glóbulos blancos ($AG = 1$); no obstante, se podría incluir la variable *diagnóstico* (AG) como una variable dicotómica (*dummy*). Previo al análisis de regresión, se examina el ajuste de los datos a uno de los modelos probabilísticos exponencial, Weibull o log-normal, mediante el método gráfico mostrado en la sección 4.3. Las gráficas se elaboraron mediante el paquete computacional R con el siguiente código y sintaxis:

```
> # Estimaci\'on de la sobrevida v\'ia Kaplan-Meier (EKM)
> t<-c(65,100,134,16,121,4,39,56,26,22,1,1,5,65,140,106,121)
> cen<-c(rep(1,14),rep(0,3))
> cgb<-c(2.3,0.75,4.3,2.6,6.0,10.5,10.0,17.0,5.4,7.0,9.4,
```

Tabla 5.1. Datos de tiempo a la muerte por leucemia

Tiempo	AG: positivo	CGB	Tiempo	AG: negativo	CGB
65	1	2.3	56	0	4.4
140^{+}	1	0.75	65	0	3.0
100	1	4.3	17	0	4.0
134	1	2.6	7	0	1.5
16	1	6.0	16	0	9.0
106^{+}	1	10.5	22	0	5.3
121	1	10.0	3	0	10.0
4	1	17.0	4	0	19.0
39	1	5.4	2	0	27.0
121^{+}	1	7.0	3	0	28.0
56	I	9.4	8	0	31.0
26	I	32.0	4	0	26.0
22	I	35.0	3	0	21.0
1	I	100.0	30	0	79.0
1	1	100.0	4	0	100.0
5	1	52.0	43	0	100.0
65	1	100.0			

Tomado de Lawless [43, p. 277].

CGB: en unidades de 1000.

```
+ 32.0,35.0,100.0,100.0,52.0,100.0)
> datos<-cbind(t,cen,cgb)
> require(survival)
> datos<-as.data.frame(datos)
> i<-order(datos$t)
> datos<-datos[i,]
> ekm<-survfit(Surv(datos$t,datos$cen)~1)
> st<-ekm$surv
> tekm<-ekm$time
> inv_Phi<-qnorm(st)
> # Gr\'aficas de probabilidad
> par(mfrow=c(1,3))
> plot(tekm, -log(st),pch=16,xlab="Tiempo",sub="Exponencial",
```

```
+ ylab="-log(S(t))")
> abline(lm(-log(st)~tekm))
> plot(log(tekm),log(-log(st)),pch=16,xlab="log(tiempo)",
+ sub="Weibull",ylab="log(-log(S(t))")
> abline(lm(log(-log(st))~log(tekm)))
> plot(log(tekm),inv_Phi,pch=16,xlab="log(tiempo)",
+ sub="log-normal",ylab=expression(Phi^-1 * (S(t))))
> abline(lm(inv_Phi~log(tekm)))
```

Por la apreciación visual del panel de gráficas contenido en la figura 5.1, se puede afirmar que los modelos que más siguen apropiadamente a los datos son el exponencial y el Weibull. No obstante, con propósitos meramente ilustrativos, se ajusta un modelo de regresión con predictor lineal simple sobre la variable *conteo de glóbulos blancos* (*CGB*) con cada uno de estos tres modelos probabilísticos.

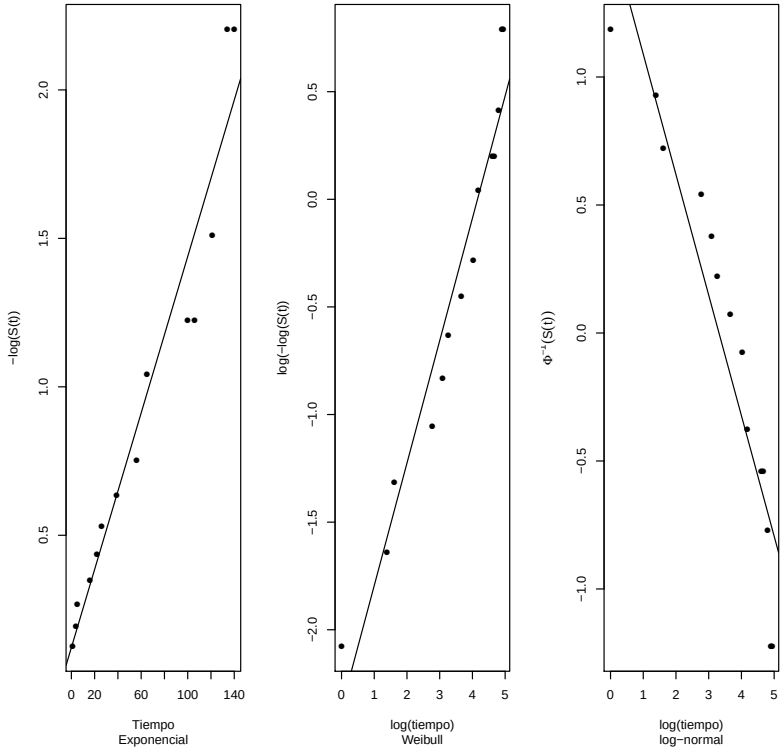

Figura 5.1. Ajuste de datos de la tabla 5.1.

El modelo de regresión de la variable respuesta *tiempo hasta la muerte* sobre la covariable *conteo de glóbulos blancos* (*CGB*), de acuerdo con (5.3.28), transformado a un modelo lineal simple, es

$$Y = \ln(T) = \beta_0 + \beta_1(CGB) + \upsilon. \tag{5.3.28}$$

La tabla 5.2 contiene las estimaciones de los parámetros asociados con el modelo de regresión ajustado y los parámetros de escala del respectivo modelo probabilístico. Las estimaciones fueron hechas mediante el paquete computacional R cuyo código se transcribe a continuación.

```
> reg_exp<-survreg(Surv(datos$t, datos$cen)~datos$cgb,
+ dist='exponential')
> reg_exp
> reg_weibull<-survreg(Surv(datos$t, datos$cen)~datos$cgb,
+ dist='weibull')
> reg_weibull
> alpha<-1/reg_weibull$scale
> alpha
> reg_lognormal<-survreg(Surv(datos$t, datos$cen)~datos$cgb,
+ dist='lognormal')
> reg_lognormal
```

Tabla 5.2. Regresión del tiempo sobre CGB

Modelo de regresión ajustado		
Exponencial	Weibull	log-normal
$\hat{\beta}_0 = 3.8187$	$\hat{\beta}_0 = 3.6933$	$\hat{\beta}_0 = 2.8945$
$\hat{\beta}_1 = 0.0172$	$\hat{\beta}_1 = 0.021$	$\hat{\beta}_1 = 0.0245$
$\alpha = 1$	$\alpha = 0.7911$	$\sigma = 1.7744$

Para decidir entre el modelo exponencial y el Weibull para estos datos, como la distribución exponencial es de la familia Weibull, se emplea la razón de verosimilitud (4.3.28), y el valor de la estadística es

$$RV = 2(71.61766 - 71.03938) = 1.1566,$$

el cual tiene asociado, para una distribución ji-cuadrado con un grado de libertad ($\chi^2_{(1)}$), un valor $p = 0.2822$. Con esto se evidencia, moderadamente,

que la distribución exponencial sigue o se ajusta a estos datos. No obstante, dado que el EMV del parámetro de escala para el modelo Weibull es $\alpha = 0.7911$, que difiere de 1.0, se podrían modelar estos datos desde una regresión Weibull como se muestra en la tabla 5.2.

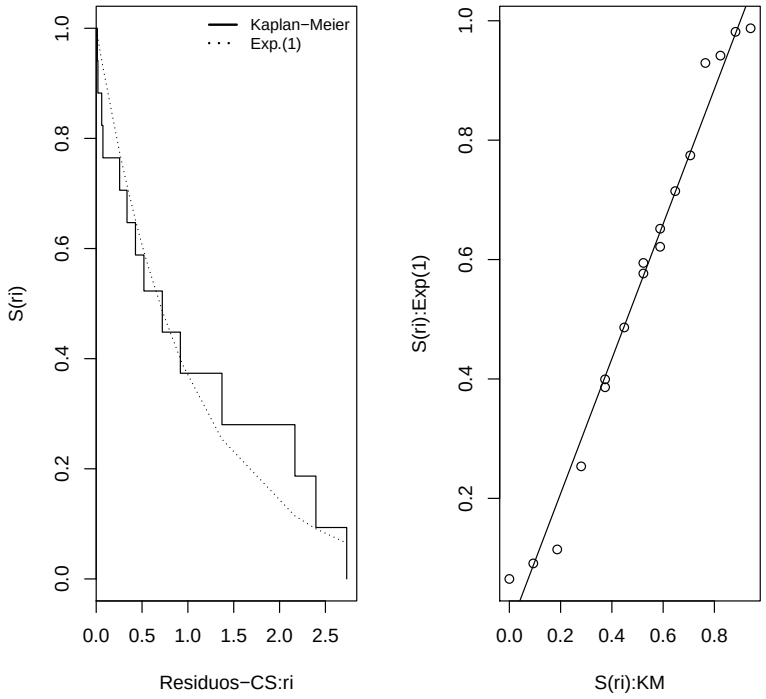

Figura 5.2. Ajuste a la exponencial de datos de sobrevida por leucemia.

Si el modelo exponencial es el adecuado para estos datos, entonces los respectivos residuos de Cox-Snell deben considerarse como una muestra aleatoria de la distribución exponencial unitaria ($\theta = 1$). Las gráficas contenidas en la figura 5.2 corroboran que la distribución exponencial se ajusta de manera moderada a estos datos. En el panel izquierdo están las funciones de sobrevida evaluadas en los residuales de Cox-Snell (r_i), estimadas mediante Kaplan-Meier, y bajo el modelo de la exponencial unitaria, respectivamente. En el panel derecho están los pares de puntos correspondientes a las estimaciones en los residuales (ordenados) obtenidos a partir del estimador de Kaplan-Meier y la exponencial unitaria, $(\hat{S}(r_i)_{KM}, \hat{S}(r_i)_{EXP})$. Si el ajuste es

adecuado, la nube de puntos debe aproximarse a una línea recta; en la figura derecha se aprecia un buen ajuste del modelo exponencial a los datos.

En el modelo de regresión exponencial se quiere verificar si la variable de conteo de glóbulos blancos en estos pacientes afecta significativamente su tiempo de sobrevida. Para este propósito se debe verificar la hipótesis $H_0 : \beta_1 = 0$. Mediante la estadística de razón de verosimilitud (4.3.22) se verifica esta hipótesis, de acuerdo con los datos y mediante el paquete computacional R se obtiene la razón de verosimilitud:

$$RV = 2(74.06643 - 71.61766) = 4.89754.$$

Esta estadística tiene asociada una distribución ji-cuadrado con un grado de libertad. El valor p correspondiente al valor de $RV = 4.89754$ es $p = 0.027$, en consecuencia, si previamente se ha asumido un nivel de significancia de $0.05 \equiv 5\,\%$, se rechaza la hipótesis nula; es decir, los datos evidencian y advierten sobre el efecto significativo del conteo de glóbulos blancos sobre los tiempos de sobrevida de estos pacientes.

La función de sobrevida para el modelo de regresión exponencial ajustado para los datos de leucemia, de acuerdo con la ecuación (5.3.1), es

$$\begin{aligned}\hat{S}(t|CGB) &= \exp\{-(\exp[3.8187 + 0.0172(CGB)]t)\} \\ &= \exp\left\{-te^{3.8187}e^{0.0172(CGB)}\right\}.\end{aligned}$$

A partir de la descomposición o factorización anterior se puede concluir que para valores grandes de CGB se necesita, para un tiempo t fijo, que la probabilidad de sobrevida estimada sea menor.

Ahora se incluye la variable AG en el modelo de regresión, de manera que se tienen los datos para los 33 pacientes como se muestran en la tabla 5.1. De esta manera, el modelo incluye las variables *conteo de glóbulos blancos* CGB (en unidades de 1000) y la variable que define los grupos positivo ($AG = 1$) o negativo ($AG = 0$).

Una ilustración, como se muestra en la figura 5.1, incluye el grupo de características negativas en las células de los glóbulos blancos ($AG = 0$), mediante un examen o prueba de laboratorio. La figura 5.3 muestra las nubes de puntos para los dos grupos; en cada panel se tienen los datos considerados provenientes de un modelo exponencial, Weibull o log-normal, respectivamente, tanto para los grupos positivo como negativo. Nuevamente, a simple vista, se advierte que para los datos que se asocian con un modelo probabilístico exponencial hay un ajuste relativamente adecuado (proximidad a una línea recta). En consecuencia, se reformula el modelo (5.3.28) con la inclusión de la covariable AG. Con estas covariables se pueden construir los

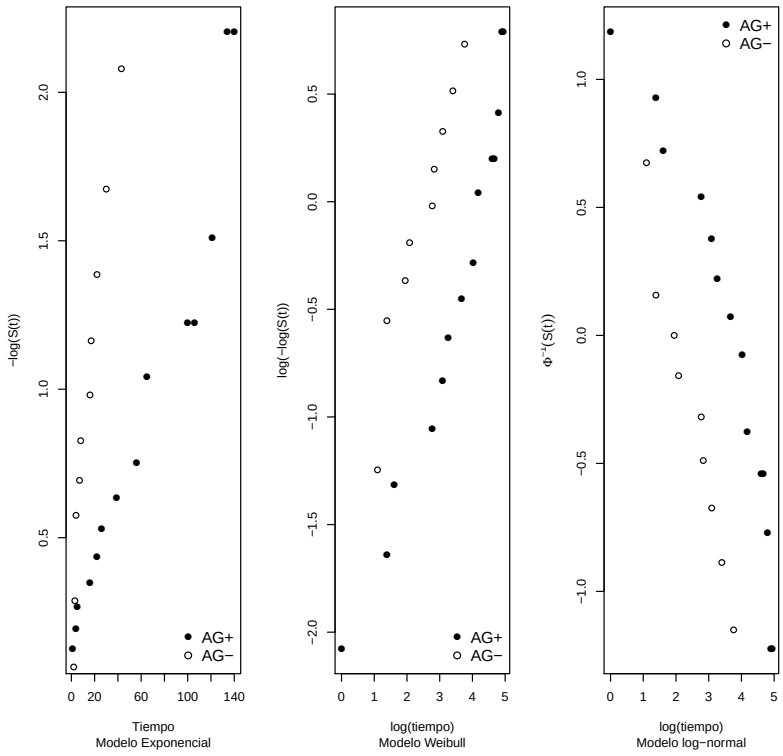

Figura 5.3. Gráficos de probabilidad para los datos de la tabla 5.1.

siguientes cinco modelos: (i) "nulo" (sin covariables), (ii) con la covariable CGB, (iii) con la covariable AG, (iv) con las covariables CGB y AG y (v) con las covariables CGB, AG y la interacción $CGB * AG$. La tabla 5.3 muestra las estimaciones y el logaritmo de la verosimilitud para cada uno de los cinco modelos.

El código R, para calcular las estimaciones y demás estadísticas asociadas con los cinco modelos que se consignan en la tabla 5.3, es el siguiente:

```
##### REGRESION EXPONENCIAL CON LAS COVARIABLES CGB, AG y CGB*AG
t<-c(65,100,134,16,121,4,39,56,26,22,1,1,5,65,140,106,121,56,65,
+17,7,16,22,3,4,2,3,8,4,3,30,4,43)
cen<-c(rep(1,14),rep(0,3),rep(1,16))
cgb<-c(2.3,0.75,4.3,2.6,6.0,10.5,10.0,17.0,5.4,7.0,9.4,32.0,
+35.0,100.0,100.0,52.0,100.0,4.4,3.0,4.0,1.5,9.0,5.3,10.0,19.
+0,27.0,28.0,31.0,26.0,21.0,79.0,100.0,100.0)
AG<-c(rep(0,17),rep(1,16))
```

```
datos<-as.data.frame(cbind(t,cen,cgb,AG))
attach(datos)
require(survival)
reg_exp1<-survreg(Surv(t, cen)~1, dist='exponential')
reg_exp1
reg_exp1$loglik
reg_exp2<-survreg(Surv(t, cen)~cgb, dist='exponential')
reg_exp2
reg_exp2$loglik
reg_exp3<-survreg(Surv(t, cen)~AG, dist='exponential')
reg_exp3
reg_exp3$loglik
reg_exp4<-survreg(Surv(t, cen)~cgb+AG, dist='exponential')
reg_exp4
reg_exp4$loglik
reg_exp5<-survreg(Surv(t, cen)~cgb+AG+cgb*AG, dist='exponential')
reg_exp5
reg_exp5$loglik
```

La verificación de la significancia de la interacción $CGB * AG$, equivalente a verificar la hipótesis $H_0 : \beta_3 = 0$, se hace mediante la razón de verosimilitud desde los datos de la tabla 5.3; esta es

$$RV = 2(135.0351 - 133.8071) = 2.456,$$

la cual tiene asociado un valor $p = P(\chi^2_{(1)} > 2.456) = 0.1171$ (ji-cuadrado con un grado de libertad). Este resultado sugiere que los datos no respaldan la inclusión del efecto conjunto o de interacción del conteo de glóbulos blancos (CGB) y su caracterización (AG).

El paso siguiente es verificar el efecto de la covariable AG condicionada a la inclusión en el modelo de la covariable CGB y, finalmente, verificar el efecto de la covariable CGB; es decir, se hace una especie de "regresión" (en inglés, *backward*). Los resultados se resumen en la tabla 5.4.

Así, el modelo propuesto incluye las covariables CGB y AG, cuyos estimadores corresponden al modelo (iv) de la tabla 5.3.

La función de sobrevida, asociada con el modelo de regresión exponencial, para un paciente con leucemia aguda es

$$\hat{S}(t|CGB, AG) = \exp\{-t \exp[4.0274 + 0.0077(CGB) - 1.3301(AG)]\}.$$

A partir de este modelo estimado, se observa que para valores grandes de CGB la probabilidad de sobrevivencia es menor. También, de manera parcial o marginal, pacientes cuyo valor de $AG = 1$ presentan probabilidades

Tabla 5.3. Estimación de regresión exponencial de los datos de leucemia

Modelo	Covariables	Estimación	log-verosimilitud
(i)	Ninguna	$\hat{\beta}_0 = 3.7758$	-143.2746
(ii)	CGB	$\hat{\beta}_0 = 3.4829$ $\hat{\beta}_1 = 0.0096$	-141.3089
(iii)	AG	$\hat{\beta}_0 = 4.2905$ $\hat{\beta}_1 = -1.4036$	-136.2567
(iv)	CGB+AG	$\hat{\beta}_0 = 4.0274$ $\hat{\beta}_1 = 0.0077$ $\hat{\beta}_2 = -1.3301$	-135.0351
(v)	CGB+AG+CGB*AG	$\hat{\beta}_0 = 3.8187$ $\hat{\beta}_1 = 0.0172$ $\hat{\beta}_2 = -0.9396$ $\hat{\beta}_3 = -0.0169$	-133.8071

Tabla 5.4. Estimación de regresión exponencial de los datos de leucemia

Efecto	Hipótesis nula	Razón de verosimilitud (Valor p)
Interacción $CGB * AG$	$H_0 : \beta_3 = 0$	$2(135.0351 - 133.8071) = 2.456$ (0.1171)
de $AG\|CGB$	$H_0 : \beta_2 = 0$	$2(141.3089 - 135.0351) = 12.5476$ (0.0004)
de CGB	$H_0 : \beta_1 = 0$	$2(143.2746 - 141.3089) = 3.9314$ (0.0474)

de sobrevivir menores que los pacientes para los cuales $AG = 0$. Las curvas de sobrevida, ajustadas por el modelo de regresión exponencial, para cua-

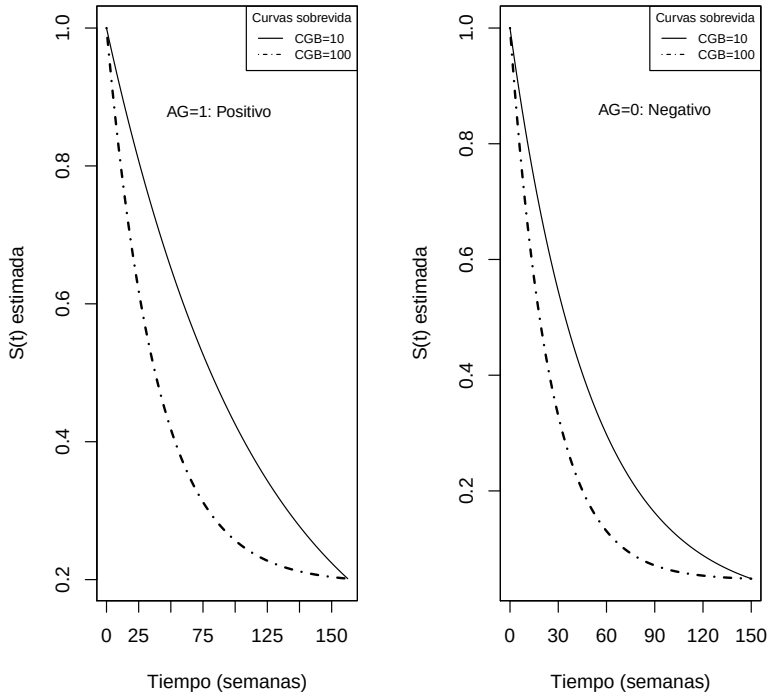

Figura 5.4. Curvas de sobrevida (leucemia) ajustadas al modelo exponencial.

tro perfiles de pacientes, se muestran en la figura 5.5. En el panel izquierdo están las gráficas de la función de sobrevida para los pacientes con leucemia cuyo examen AG es positivo ($AG = 1$), mientras que en el panel de la derecha están las funciones de sobrevida para pacientes con examen AG negativo ($AG = 0$); en cada panel se grafican las curvas de sobrevida para los pacientes con un conteo de glóbulos blancos de 10 y 100 (en escala de 1000), respectivamente. Se puede concluir que los pacientes con menor número en el conteo de glóbulos blancos tienen una mayor probabilidad de sobrevivir que quienes tienen un valor mayor en esta variable. Esta conclusión es válida para los dos pronósticos, AG positivo y AG negativo.

En la figura 5.5 se muestran cuatro curvas de sobrevida. Dentro del grupo de pacientes con examen AG positivo se consideran el subgrupo de quienes tienen un CGB igual a 10 y quienes tienen un CGB de 100; las mismas curvas se tienen para el grupo AG negativo. Se puede observar en el or-

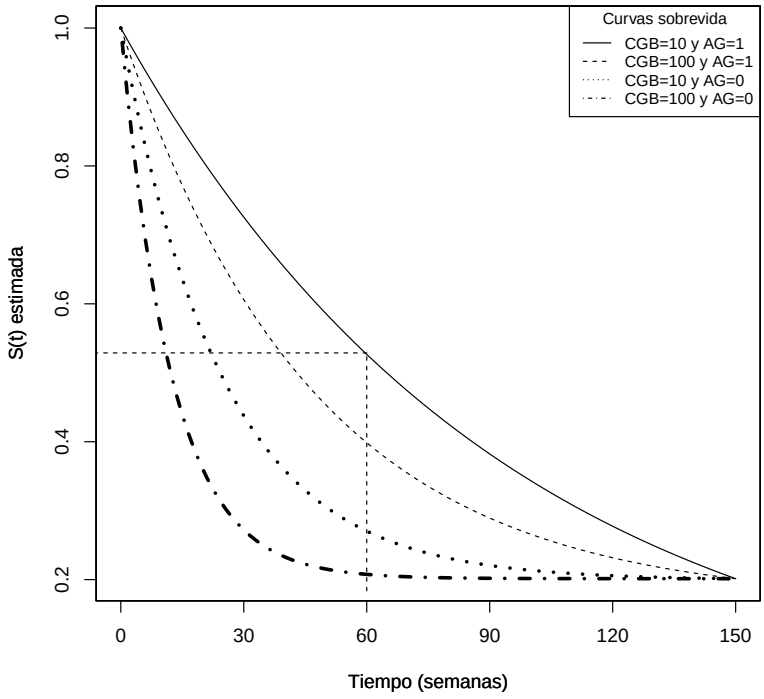

Figura 5.5. Comparación de curvas de sobrevida (leucemia) ajustadas al modelo de regresión exponencial.

den decreciente de las curvas, en términos de la probabilidad de sobrevida, que el grupo de los diagnosticados positivamente es mayor que el grupo de los diagnosticados negativamente. Dentro de cada grupo el orden es el mismo, la mayor probabilidad de sobrevida la tienen quienes muestran un conteo de glóbulos blancos menor. Se resalta que para un tiempo $t = 60$ semanas, el grupo con mayor probabilidad de sobrevivir a este tiempo es el que tienen un AG positivo ($AG = 1$) y un conteo de glóbulos blancos de 10 ($CBB = 10$), mientras que el grupo con el pronóstico de sobrevida más pobre es el grupo con AG negativo y un conteo de glóbulos blancos de 100.

Capítulo

seis

Modelo de riesgos proporcionales de Cox

$$\ln\left(\frac{h(t|x)}{h_0(t)}\right) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

6.1. Introducción

Como en el capítulo anterior, se trata de modelar la variable *tiempo* en función de algunos atributos o factores relacionados con la unidad (individuo u objeto), llamados covariables. El modelamiento se puede hacer sobre la función de sobrevida o sobre la función de riesgo; en este capítulo se modela la función de riesgo. Así, se trata la función de riesgo dependiente del tiempo para el evento y de un conjunto de covariables.

Un punto crítico para usar un modelo de regresión paramétrica en el análisis de tiempo hasta el evento es la dificultad para determinar la validez de un modelo probabilístico subyacente a la distribución de la variable *tiempo*. Incluso, la selección de una distribución paramétrica puede no reflejar necesariamente las verdaderas trayectorias individuales del proceso del tiempo hasta el evento; por lo tanto, las estimaciones de los parámetros, derivadas de un modelo de regresión asociado con una distribución de la variable *tiempo* mal especificada, pueden ser sesgadas.

Además, muchos investigadores están más interesados en observar los efectos de algunas covariables sobre el riesgo de una ocurrencia del evento de interés que en el forma de una distribución de tiempo específica. Si bien el modelo de riesgo proporcional es altamente acogido con satisfacción en muchas disciplinas aplicadas, el uso de modelos de regresión paramétrica es a veces inconveniente porque solo las funciones de distribución exponencial y Weibull pueden ser formuladas como un modelo de riesgo proporcional. Resulta necesario estructurar un modelo de regresión con el cual las estimaciones de los efectos de las covariables sobre la función de riesgo no requieran de la especificación de una función de distribución asociada con el tiempo al evento.

El interés se centra en explicar y pronosticar el riesgo a la ocurrencia del evento en función de las covariables y la variable *tiempo hasta el evento*. El procedimiento de modelamiento es similar al de los modelos de regresión, pues tiene los componentes de estructuración del modelo, la estimación de los parámetros y el diagnóstico de la calidad y sensibilidad del modelo; esto en términos de los supuestos estadísticos y probabilístico requeridos y de la información contenida en los datos disponibles.

Si todas las unidades están sujetas a una línea base común de la distribución del tiempo, y las diferencias en la tasa de riesgo instantáneo se pueden atribuir y reflejar principalmente a través del efecto multiplicativo de las covariables, la distribución de los tiempos sería una constante y, por lo tanto, podría "separarse" potencialmente desde una función de verosimilitud específica.

El modelo que recoge estas consideraciones es el *modelo de riesgos proporcionales*. Fue propuesto por Cox [14] y por esta razón se le denomina el *modelo de riesgos proporcionales de Cox*. Este modelo no requiere del supuesto de algún modelo probabilístico particular para la variable *tiempo hasta el evento*.

En este capítulo se muestra la estructura del modelo de Cox, el proceso de estimación de los parámetros que identifican al modelo, las técnicas de validación del modelo y también se tratan algunos tópicos especiales como las covariables tiempo-dependientes y los modelos estratificados.

6.2. Modelo de Cox

Se supone que cada unidad u observación bajo estudio esta sujeta a una tasa de riesgo instantáneo, $h(t)$, de experimentar un evento en particular, donde $t > 0$. De manera similar al modelo paramétrico de riesgo proporcional, el efecto de las covariables en el modelo de Cox se especifica mediante un término multiplicativo $\exp(\mathbf{x}'\beta)$, el cual, por la naturaleza de la función de riesgo, es no negativo. La ecuación básica del modelo de Cox es exactamente la misma ecuación (5.2.1), y esta es dada por

$$h(t \mid \mathbf{x}) = h_0(t) \exp(\mathbf{x}'\beta), \tag{6.2.1}$$

donde $h_0(t)$ representa una función de riesgo base, la cual es una línea de base arbitraria y no especificada dependiente solo del tiempo. El vector $\beta = (\beta_1, \ldots, \beta_p)$ contiene los parámetros asociados con los efectos de las covariables, $\mathbf{x} = (x_1, x_2, \ldots, x_p)$, sobre la tasa de riesgo; este vector tiene el mismo tamaño que el vector de covariables x. Se asume que el efecto de las x sobre la tasa de riesgo $h(\cdot)$ es multiplicativo, de manera que el valor pronosticado para h al tiempo t dado $\mathbf{x}$, denotado por $h(t|\mathbf{x})$, varía en los números reales positivos, es decir, en el rango $(0, \infty)$.

El modelo (6.2.1) está definido por dos componentes, β y $h_0(t)$. De manera equivalente se puede definir el modelo de Cox a través de la función de sobrevida. Así, por la expresión (1.5.4), $S(t) = e^{-\int_0^t h(u)du} = e^{-H(t)}$, luego, $S_0(t) = e^{-H_0(t)}$. La función de sobrevida para $\mathbf{T}$, dadas las covariables $\mathbf{x}$, es

$$\begin{aligned} S(t|\mathbf{x}) &= \exp\left\{-\int_0^t h(u)du\right\} \\ &= \exp\left\{-\int_0^t h_0(u)\exp(\mathbf{x}'\beta)du\right\} \\ &= \exp\left\{-\exp(\mathbf{x}'\beta)\int_0^t h_0(u)du\right\} \\ &= \left[\exp\left\{-\int_0^t h_0(u)du\right\}\right]^{\exp(\mathbf{x}'\beta)} \\ &= [S_0(t)]^{\exp(\mathbf{x}'\beta)}. \end{aligned} \tag{6.2.2}$$

Dado que el modelo (6.2.1) puede ser expresado en la forma

$$\ln\left(\frac{h(t|\mathbf{x})}{h_0(t)}\right) = \eta = \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_p x_p, \tag{6.2.3}$$

el modelo de riesgos proporcionales se considera como un modelo lineal para el logaritmo de la razón de verosimilitud. Por la expresión (6.2.3), al modelo de Cox se le denomina *modelo log-aditivo*.

Note que en el componente lineal no hay término constante; en caso de que lo hubiera, este actuaría de manera multiplicativa sobre $h_0(t)$ y podría incluirse en este, pues no depende de ninguna covariable.

Al modelo de Cox (6.2.1) también se le denomina modelo *semiparamétrico*, porque la forma paramétrica está en el efecto de las covariables (β's) del componente $\exp(\mathbf{x}'\beta)$, mientras que la linea de riesgo base se considera no paramétrica. Esto se interpreta en dos sentidos; uno es que no tiene efecto de covariables, es decir, no tiene parámetros, y el otro es que no se asume la distribución particular para la variable *tiempo hasta el evento*. En resumen, el componente $h_0(t)$ no tiene parámetros, mientras que el componente $\exp(\mathbf{x}'\beta)$ sí los tiene; de ahí, el término *semiparamétrico*.

6.2.1. Variables del componente lineal

El componente lineal, $\eta = \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_p x_p$, del modelo de Cox puede estar conformado por dos tipos de covariables; unas se denominan, en este texto, *variables*, y las otras se llaman *factores*. Una *variable* es aquella que toma valores en una escala al menos de intervalo (continua), tales como la edad, el peso y la temperatura. Un *factor* es una variable que toma un número finito de valores, los cuales se denominan *niveles*. Por ejemplo, el sexo biológico es un factor con dos niveles, la raza de una persona o un animal es un factor

con varios niveles, la marca de un producto es un factor con un nivel por marca específica; así, cada factor induce un subgrupo (estrato) de unidades.

Los modelos que contienen términos correspondientes a factores pueden expresarse como combinaciones lineales de los niveles que conforman un factor. Para ello, se definen unas variables indicadoras o "ficticias" (en inglés, *dummy*) para los niveles de cada factor. Es importante saber cuál es la codificación que cada paquete estadístico emplea para cada factor. En el caso de una variable dicotómica, X, los valores que toma esta variable son $X = 1$ para una característica y $X = 0$ para la característica complementaria. En general, si una covariable es un factor con a niveles, se generan $a - 1$ variables de codificación $Z_2, Z_3, \ldots Z_a$, como se muestra en la tabla 6.1. Así, la mayoría de los paquetes, por defecto, tienen incorporada esta

Tabla 6.1. Codificación de los niveles de un factor

Nivel del factor	Variables "ficticias"			
k	Z_2	Z_3	$\cdots$	Z_a
1	0	0	$\cdots$	0
2	1	0	$\cdots$	0
3	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
a	0	0	$\cdots$	1

codificación, a menos que el usuario quiera introducir otra. Con una codificación como la mostrada en la tabla 6.1, los paquetes reportan estimaciones de los parámetros asociadas con las variables Z_2 a Z_a, de manera que si se quiere tener un pronóstico del primer nivel, se deben tomar los coeficientes asociados con las variables ficticias Z_2 a Z_a iguales a cero.

Aunque en la presentación del modelo de Cox no se hace explícito, las variables también pueden considerarse como *tiempo-dependientes*. De esta manera, el modelo de Cox se debe expresar como

$$h(t \mid \mathbf{x}(\mathbf{t})) = h_0(t) \exp(\mathbf{x}(\mathbf{t})'\beta), \tag{6.2.4}$$

donde $\mathbf{x}(\mathbf{t}) = (x_1(t), x_2(t), \ldots, x_p(t))'$. Las variables tiempo-dependientes o dinámicas son variables que toman valores dependientes del tiempo de observación, por ejemplo, la dosis suministrada a un paciente no es la misma todo el tiempo, pues esta puede variar de acuerdo con la evolución de su enfermedad. Otro ejemplo tiene que ver con las calidades del lubricante de

una máquina. Características tales como densidad, viscosidad y temperatura son dependientes del número de unidades de tiempo que este lleva en uso dentro de la máquina. En el caso de que las variables sean independientes del tiempo, $x_j(t) = x_j$, para $j = 1, 2, \ldots, p$, como se asume en el modelo (6.2.1).

Un aspecto que refleja el predictor lineal $\eta = \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_p x_p$ es que cada variable, x_j, aparece solo una vez y de manera aditiva con el correspondiente parámetro β_j, para $j = 1, 2, \ldots, p$. Suele presentarse que las variables de manera individual y parcial o marginal tienen un efecto sobre la función de riesgo, mientras que su efecto conjunto o combinado es diferente. Por ejemplo, la temperatura y la humedad pueden tener un efecto en el riesgo de que una máquina presente una falla, así como pasa, en el caso de un enfermo, con la condición de ser hipertenso o no y con el tipo de fármaco que se le suministra para la enfermedad; en ambos casos el riesgo puede ser diferente si se consideran conjuntamente estas dos variables. Esto se conoce como la *interacción*. Generalmente, la interacción se ha considerado como la combinación de los niveles de varios factores. En el caso de la temperatura y la humedad, si estas se consideran en los niveles alto y bajo, se tendrían cuatro factores posiblemente interactuando sobre la variable respuesta. Para el caso de dos covariables, el predictor lineal de un modelo con interacción es de la forma

$$\eta = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2,$$

donde el parámetro β_3 refleja el efecto de la interacción entre la covariable x_1 y la covariable x_2; sean estas de tipo continuo o categóricas (factores).

En general, en el predictor lineal pueden considerarse variables, factores e interacciones de manera simultánea. El problema está en la información disponible para la estimación de los parámetros y de la capacidad de interpretación de los componentes y resultados del modelo. Las variables categóricas o factores se pueden introducir al modelo como variables ficticias (*dummy*).

El modelo de Cox se denomina *de riesgos proporcionales* porque la razón de las tasas de falla de dos unidades diferentes es constante en el tiempo. Si las dos unidades, j y k, se definen por el valor de sus vectores de covariables x_j y x_k, la razón de sus tasas de falla es

$$\frac{h(t|x_j)}{h(t|x_k)} = \frac{h_0(t)\exp\left\{x_j'\beta\right\}}{h_0(t)\exp\left\{x_k'\beta\right\}} = \exp\left\{(x_j' - x_k')\beta\right\},$$

la cual no depende del tiempo sino de la diferencia entre los perfiles de las unidades $(x_j' - x_k')$. De manera que, si una unidad tiene un riesgo al evento

el doble que el de otra unidad, esta razón se mantiene en todo el tiempo de seguimiento. Este es uno de los supuestos estructurales del modelo de regresión de Cox, es decir que la tasa del evento se mantiene proporcional o, equivalentemente, que las tasas del evento acumuladas son también proporcionales.

El supuesto de riesgos proporcionales tiene la consecuencia de que las funciones de supervivencia para valores distintos de variables explicativas $\mathbf{x}_1$ y $\mathbf{x}_2$ nunca se cruzan. Por lo tanto, $S(t|\mathbf{x}_1) > S(t|\mathbf{x}_2)$ ($S(t|\mathbf{x}_1) < S(t|\mathbf{x}_2)$) se cumple para todos los valores de t, esto indica que uno de los valores, $\mathbf{x}_1$ o $\mathbf{x}_2$, ofrece mayores probabilidades de sobrevida en todo el tiempo.

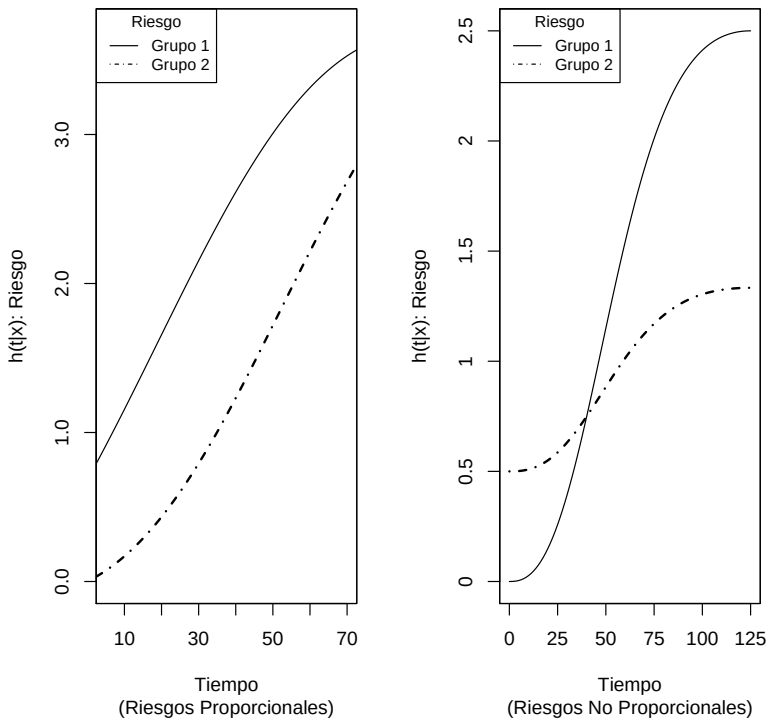

Figura 6.1. Gráfica de riesgos proporcionales y riesgos no proporcionales.

Una herramienta para detectar si se cumple o no el supuesto de proporcionalidad sobre un conjunto de datos son las gráficas de las funciones de riesgo empíricas o estimadas. Por ejemplo, si para diferentes valores de una variable categórica, las gráficas de las funciones de riesgo de cada nivel de la variable son aproximadamente paralelas, se puede considerar que el su-

puesto de proporcionalidad se cumple. La figura 6.1 contiene dos escenarios hipotéticos; en cada panel se ha trazado la curva asociada con la función de riesgo para los dos grupos determinados por una variable categórica o un factor (grupo 1 y grupo 2). En el panel izquierdo se observa que el supuesto de proporcionalidad se tiene, mientras que el panel derecho se muestra el incumplimiento de la proporcionalidad.

Dada una muestra aleatoria de tiempos hasta el evento posiblemente censurados, (t_i, δ_i), $i = 1, 2, \ldots, n$, y su correspondiente vector de covariables $\mathbf{x}_i$, se pretende estimar β y $h_0(t)$ (y, por lo tanto, $S_0(t)$). En Cox [14, 15], se propone y se formaliza el método de estimación conocido como *método de la verosimilitud parcial*, el cual se describe en la siguiente sección.

6.3. Estimación en el modelo de Cox

Se asume que se tienen n observaciones $t_1, t_2, \ldots, t_n$ de tiempo hasta el evento, sin empates, posiblemente censuradas a derecha por las constantes $c_1, c_2, \ldots, c_n$, y sea z_i mín$\{t_i, c_i\}$, para $i = 1, 2, \ldots, n$. Algunas observaciones serán censuradas, $z_i = c_i$ $(\delta_i = 0)$, y algunas no serán censuradas, es decir, $z_i = t_i$ $(\delta_i = 1)$. En particular, asuma que hay k tiempos no censurados y, en consecuencia, $n - k$ tiempos censurados.

El modelo de regresión de Cox, además de la función línea base $h_0(t)$, se define por los coeficientes contenidos en el vector β, que caracterizan y cuantifican el efecto de las covariables sobre la tasa de riesgo hasta el evento. Estos parámetros deben estimarse a través de los datos muestrales.

El procedimiento de estimación e inferencia sobre los parámetros es el método de máxima verosimilitud presentado en la sección 4.2 para modelos paramétricos.

No obstante, la presencia del componente $h_0(t)$ en la función de verosimilitud hace que este método no resulte apropiado para estimar los parámetros del componente lineal. A continuación se muestra la inconveniencia del método de máxima verosimilitud seguido en la sección 4.2. En este sentido, la función de verosimilitud vendría dada por

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} [f(t_i|\mathbf{x})]^{\delta_i} [S(t_i|\mathbf{x})]^{1-\delta_i} \\ &= \prod_{i=1}^{n} [h(t_i|\mathbf{x})]^{\delta_i} S(t_i|\mathbf{x}). \end{aligned} \tag{6.3.1}$$

Para el modelo de Cox, por (6.2.2), la función de verosimilitud es

$$L(\beta) = \prod_{i=1}^{n} \left[h_0(t_i) \exp\left\{\mathbf{x}_i'\beta\right\}\right]^{\delta_i} \left[S_0(t_i)\right]^{\exp(\mathbf{x}_i'\beta)}, \tag{6.3.2}$$

la cual es una función del componente no paramétrico $h_0(t)$; la acepción "no paramétrica" tiene dos sentidos válidos. Por una parte, no tiene de manera explícita parámetros y, por otra parte, para la variable *tiempo* no se asume distribución alguna.

Una salida a la dificultad de conocer la forma de la función $h_0(t)$ consiste en condicionar la verosimilitud de manera que "elimine" o suprima esta función.

Como el procedimiento para estimar β no depende de $h_0(t)$, Cox [14] considera que los tiempos censurados no contribuyen con información en la estimación de β, dado que $h_0(t)$ se podría considerar igual a 0 en los intervalos intra-tiempos observados. La censura de una unidad tiene el efecto simple de que hay una unidad menos en riesgo de experimentar el evento al momento de la censura, de manera que las unidades censuradas se excluyen de los conjuntos de unidades expuestas al evento en el transcurso del tiempo. En consecuencia, como los tiempos de censura se mantendrán fuera de los cálculos directos en la estimación de β, entonces se consideran y ordenan los k tiempos observados (no empatados) como

$$t_{(1)} < t_{(2)} <, \cdots, < t_{(k)}.$$

Se emplean estos tiempos de forma directa en la construcción de la verosimilitud (verosimilitud condicional o parcial) formalizada en Cox [15]. A continuación, se sigue la presentación de Smith [64] y Collet [10]. Sea $\mathbf{x}_{(i)}$ el vector de covariables asociado con el tiempo $t_{(i)}$. Los $n-k$ tiempos censurados entran a la verosimilitud a través de su pertenencia a los k conjuntos de riesgo, $\mathcal{R}_j$, con $j = 1, 2, \ldots, k$; es decir, el conjunto de las unidades expuestas al evento al momento $t_{(j)}^{-}$. El exponente "-" significa "inmediatamente" antes que $t_{(j)}$.

Se emplea el concepto de conjunto expuesto al evento (conjunto en riesgo) directamente en la construcción de la verosimilitud condicional. Se retoma la sección 1.5 para la construcción de la función de riesgo, esta vez condicionada al vector de covariables $\mathbf{x}$. Se puede escribir la función de riesgo condicionada a las covariables como

$$h(t|\mathbf{x}) \approx \frac{P(tT@ < \mathbf{T_x} < \mathbf{t} + \Delta\mathbf{t}|\mathbf{T} > \mathbf{t})}{\Delta t}, \tag{6.3.3}$$

la cual se puede escribir como

$$h(t|\mathbf{x})\Delta t \approx P(tT@ < \mathbf{T_x} < \mathbf{t} + \Delta\mathbf{t}|\mathbf{T} > \mathbf{t}) \tag{6.3.4}$$

$$\approx P(\text{Evento al momento } t|\text{No evento hasta } t). \tag{6.3.5}$$

Se reitera que en el procedimiento de estimación del modelo de Cox se consideran los tiempos observados (no censurados) junto con sus conjuntos de riesgo asociados.

A manera de ejemplo, considere las 6 unidades con sus respectivos tiempos hasta el evento como se muestran en la tabla 6.2. Para estos datos, $n = 6$,

Tabla 6.2. Tiempos al evento

Unidad	Tiempo	Censura
i	z_i	δ_i
1	7	1
2	124	1
3	88	0
4	13	1
5	2	0
6	79	1

Tomado de Smith [64, p. 169].

$k = 4$, y los datos observados (no censurados) ordenados son $t_{(1)} = 7$, $t_{(2)} = 13$, $t_{(3)} = 79$, y $t_{(4)} = 124$. Los conjuntos de unidades expuestas al evento en cada uno de estos tiempos son:

$$\mathcal{R}_1 = \{1, 2, 3, 4, 6\},\ \mathcal{R}_2 = \{2, 3, 4, 6\},\ \mathcal{R}_3 = \{2, 3, 6\},\ \mathcal{R}_4 = \{2\}.$$

Ahora se muestra la construcción de la verosimilitud. Considere el tiempo $t_{(j)}$ con su respectivo conjunto de riesgo al evento $\mathcal{R}_j$, con $j = 1, 2, \ldots, k$.

La probabilidad de que sobre una unidad de $\mathcal{R}_j$ con covariables $\mathbf{x}_j$ ocurra el evento al momento $t_{(j)}$, dado que a una unidad del conjunto $\mathcal{R}_j$ le ocurre el evento en el momento $t_{(j)}$, de acuerdo con la aproximación (6.3.4), es

dada por:

$$
\begin{aligned}
&= \frac{P(\text{La unidad de } \mathcal{R}_j \text{ de covariables } \mathbf{x}_j \text{ tenga el evento en } t_{(j)})}{P(\text{Una unidad de } \mathcal{R}_j \text{ tenga el evento en } t_{(j)})} \\
&\approx \frac{h(t_{(j)}|\mathbf{x}_{(j)})\Delta t}{\sum_{l\in\mathcal{R}_j} h(t|\mathbf{x}_{(j)})\Delta t} \\
&= \frac{h(t_{(j)}|\mathbf{x}_{(j)})}{\sum_{l\in\mathcal{R}_j} h(t|\mathbf{x}_{(j)})} \\
&= \frac{h_0(t_{(j)})e^{\mathbf{x}'_{(j)}\beta}}{\sum_{l\in\mathcal{R}_j} h_0(t_{(j)})e^{\mathbf{x}'_{(j)}\beta}} \\
&= \frac{e^{\mathbf{x}'_{(j)}\beta}}{\sum_{l\in\mathcal{R}_j} e^{\mathbf{x}'_{(j)}\beta}}.
\end{aligned}
$$

Note que la función de riesgo base, $h_0(t_{(j)})$, se cancela del numerador y denominador; esto se hace porque el proceso de estimación de los β's se puede desarrollar sin el conocimiento de esta función.

Tomando esta aproximación para todos los tiempos observados e independientes $t_{(j)}$, se tiene la *verosimilitud condicional*

$$
L_c(\beta) = \prod_{j=1}^{k} \frac{\exp\left\{\mathbf{x}'_{(j)}\beta\right\}}{\sum_{l\in\mathcal{R}_j}\exp\left\{\mathbf{x}'_{(l)}\beta\right\}} = \prod_{i=1}^{n}\left(\frac{\exp\left\{\mathbf{x}'_i\beta\right\}}{\sum_{l\in\mathcal{R}_i}\exp\left\{\mathbf{x}'_{(l)}\beta\right\}}\right)^{\delta_i}, \qquad (6.3.6)
$$

donde δ_i es un indicador de ocurrencia del evento (no censura). Esta no es una función de verosimilitud en sentido estricto, ya que los datos de tiempo para el evento, tanto censurados como no censurados, no están involucrados numéricamente en el cálculo. Los datos censurados solo participan de los conjuntos de unidades expuestas a la ocurrencia del evento, $\mathcal{R}_j$, en el denominador de (6.3.6). Por esta razón se ha acuñado el término *verosimilitud parcial de Cox* sobre la expresión (6.3.6) para señalar la participación de los tiempos observados en la construcción de L_c, propuesta en Cox [14] y sustentada en el también magistral artículo [15].

Para ilustrar la metodología de la verosimilitud parcial se sigue el ejemplo de Collet [10, p. 66], quien considera el caso de 5 unidades, U_i, con $i = 1, 2, 3, 4, 5$, como se muestra en la figura 6.2.

Se observa que las unidades U_2 y U_5 tienen tiempos censurados. Los tiempos ordenados de las otras tres unidades son $t_{(1)} < t_{(2)} < t_{(3)}$, donde $t_{(1)}$ corresponde a la unidad U_3, $t_{(2)}$ a la unidad U_1 y $t_{(3)}$ a la unidad U_4.

Los conjuntos expuestos al evento inmediatamente antes de los tres tiempos ordenados se conforman por las unidades a las que no les ha ocurrido

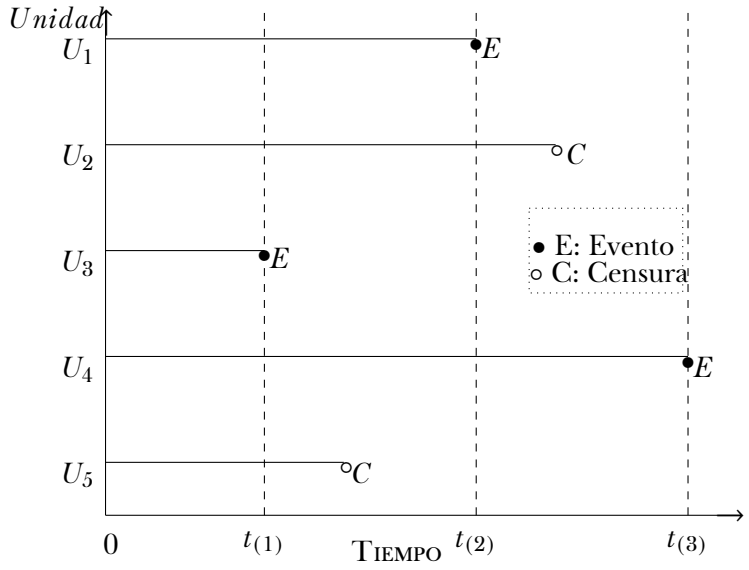

Figura 6.2. Tiempo al evento en cinco unidades.

el evento inmediatamente antes de cada uno de estos tres tiempos. Así, los conjuntos son $\mathcal{R}_1 = \{U_1, U_2, U_3, U_4, U_5\}$, $\mathcal{R}_2 = \{U_1, U_2, U_4\}$ y $\mathcal{R}_3 = \{U_4\}$.

Los perfiles, en términos de las covariables, que conforman la verosimilitud parcial (6.3.6) son $\psi(\mathbf{x}_i) = \exp\{\mathbf{x}_i'\beta\}$, para $i = 1, 2, 3, 4, 5$. Los tiempos observados no censurados son $t_{(1)}$, $t_{(2)}$ y $t_{(3)}$, donde $t_{(1)}$ corresponde a las unidades U_3, U_1 y U_4, respectivamente. Por lo tanto, los numeradores de la verosimilitud parcial de (6.3.6) son $\psi(3)$, $\psi(1)$ y $\psi(4)$, puesto que a las unidades U_3, U_1 y U_4, respectivamente, les ocurre el evento en los tres tiempos ordenados. La función de verosimilitud parcial sobre los tres tiempos es

$$\frac{\psi(3)}{\psi(1)\psi(2)\psi(3)\psi(4)\psi(5)} \times \frac{\psi(1)}{\psi(1)\psi(2)\psi(4)} \times \frac{\psi(4)}{\psi(4)}.$$

El logaritmo de la verosimilitud parcial (6.3.6) está dado por

$$l(\beta) = \ln L(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \mathbf{x}_i'\beta - \ln \sum_{l \in \mathcal{R}_i} \exp\left(\mathbf{x}_{(l)}'\beta\right) \right\}. \tag{6.3.7}$$

Los valores de β que maximizan la función de verosimilitud parcial, $l(\beta)$, se obtienen resolviendo el sistema de ecuaciones de las derivadas parciales

de cada componente de β sobre la función $l(\beta)$ igualados a cero. Esta es

$$U(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} \delta_i \left[\mathbf{x}_i - \frac{\sum_{l \in \mathcal{R}_i} \mathbf{x}_{(l)} \exp\left(\mathbf{x}'_{(l)}\beta\right)}{\sum_{l \in \mathcal{R}_i} \exp\left(\mathbf{x}'_{(l)}\beta\right)} \right] = 0, \qquad (6.3.8)$$

así, las p-ecuaciones de máxima verosimilitud son

$$U(\beta) = (\partial l(\beta)/\partial \beta_1, \partial l(\beta)/\partial \beta_2, \ldots, \partial l(\beta)/\partial \beta_p)' = \mathbf{0}_{p\times 1}. \qquad (6.3.9)$$

El sistema de las ecuaciones de máxima verosimilitud, $U(\beta) = 0$, se puede resolver mediante el método de Newton-Raphson (4.2.8), como se muestra adelante, cuya solución corresponde al estimador máximo verosímil de β; este es $\hat{\beta}_{MV}$. Este algoritmo numérico está incorporado en los paquetes estadísticos R, SAS, SPSS y MINITAB entre otros, con los cuales se calculan los errores estándar, los intervalos o las regiones de confianza y se desarrollan estadísticas de verificación de hipótesis basadas en la aproximación a la distribución normal de $\hat{\beta}_{MV}$, es decir,

$$\hat{\beta}_{MV} \xrightarrow{D} N(\beta, \mathcal{I}(\beta)^{-1}).$$

A continuación, se presenta el caso de datos empatados y su consideración, tratamiento e inclusión en el modelo de Cox.

6.3.1. Tratamiento de empates

Se entiende por *empate* a los tiempos al evento iguales en dos o más unidades. Por el supuesto de continuidad de la variable aleatoria *tiempo al evento*, **T**, no deberían presentarse empates; de otra manera, estos ocurrirían con probabilidad cero. No obstante, por circunstancias de muestreo y de redondeo o por aproximación en el registro del tiempo, pueden tenerse empates. Así, por ejemplo, si el tiempo es medido en semanas, el evento puede ocurrir sobre varias unidades en diferentes días de la misma semana; en consecuencia, como la unidad de tiempo es la semana, todas estas unidades tendrán un empate en el tiempo al evento. Lo mismo puede presentarse para el caso de medir el tiempo en meses, pues las observaciones que se presenten antes o después de la quincena serían respectivamente empatadas por aproximación por defecto o por exceso.

Los empates también pueden presentarse en observaciones censuradas, por ejemplo, en presencia de censura a derecha, los tiempos de las unidades a las que posiblemente les ocurra el evento después de cierto tiempo se pueden considerar como empatados. Así, se deben considerar e incluir las

observaciones empatadas en el cómputo de la verosimilitud parcial (6.3.6). La aproximación a la función verosimilitud, introducida por Breslow [5], se presenta a continuación.

Para acomodar las observaciones empatadas a la verosimilitud parcial, es necesario introducir la notación que se describe a continuación.

- S_j es el vector de sumas de cada una de las p-covariables para las unidades sobre las cuales ocurre el evento en el j-ésimo tiempo $t_{(j)}$, con $j = 1, 2, \ldots, r$; es decir, hay r tiempos donde ocurren empates,
- d_j es el número de unidades a las que les acontece el evento en el momento $t_{(j)}$ (número de empates),
- el h-ésimo elemento de S_j es $S_{h_j} = \sum_{k=1}^{d_j} x_{hjk}$, donde x_{hjk} es el valor de la h-ésimo variable explicativa (covariable) $h = 1, 2, \ldots, p$ para el k-ésimo de las d_j unidades $k = 1, 2, \ldots, d_j$ que tienen el evento en el j-ésimo tiempo $j = 1, 2, \ldots, r$.

Breslow [5] propone la siguiente aproximación para la función de verosimilitud parcial:

$$\prod_{j=1}^{r} \frac{\exp(\beta' S_j)}{\left[\sum_{l \in R(t_{(j)})} \exp\{\beta' X_l\}\right]^{d_j}}. \tag{6.3.10}$$

En esta aproximación, los d_j eventos que ocurren en el tipo $t_{(j)}$ son considerados distintos y ocurren secuencialmente.

La solución del sistema de ecuaciones (6.3.9) puede encontrarse mediante el método de Newton-Raphson como se describe a continuación.

Sean $U(\beta) = \frac{\partial l(\beta)}{\partial \beta}$, con $l(\beta) = \ln L(\beta)$, el vector de puntuaciones y $\mathcal{I}O(\beta) = \left(-\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right)$ la matriz de información observada. Una aproximación a la solución del sistema de ecuaciones (6.3.9) en el paso $m + 1$ es

$$\widetilde{\beta}^{(m+1)} = \widetilde{\beta}^{(m)} + \mathcal{I}O^{-1}\left(\widetilde{\beta}^{(m)}\right) U\widetilde{\beta}^{(m)}.$$

El valor de m tal que $|\widetilde{\beta}^{(m+1)} - \widetilde{\beta}^{(m)}| < \epsilon$, para algún $\epsilon > 0$, se considera que $\widetilde{\beta}^{(m+1)}$ es el estimador de máxima verosimilitud de β; se nota por $\widehat{\beta}^{(m+1)}$.

La raíz cuadrada de los elementos de la diagonal de la matriz de observación, $\mathcal{I}O^{-1}\widehat{\beta}$, corresponde a los errores estándar (*ee*) de los estimadores $(\hat{\beta}_1, \ldots, \hat{\beta}_p)$.

6.3.2. Interpretación de los coeficientes β's

A partir del modelo $h(t|\mathbf{x}) = h_0(t)e^{\mathbf{x}'\beta}$, se puede observar que el efecto de las covariables contenidas en $\mathbf{x}$ es acelerar o decelerar la función de riesgo. Por la propiedad de riesgos proporcionales del modelo se pueden interpretar los coeficientes estimados. Asuma que se tienen dos unidades i e i', las cuales tienen los mismos valores en todas las covariables, excepto en la covariable j-ésima. Se tiene que

$$\frac{h(t|\mathbf{x}_i)}{h(t|\mathbf{x}_{i'})} = \exp\left[\beta_j(x_{ij} - x_{i'j})\right],$$

lo que representa la razón de riesgos en el momento t, la cual se mantiene constante para cualquier momento t. Para especificar un poco más, suponga que la covariable x_j es una variable de pertenencia a un grupo, es decir, $x_j = 1$ significa pertenencia al grupo, mientras que $x_j = 0$ significa no pertenencia a ese grupo. El riesgo al evento de la unidades del grupo es $e^{\hat{\beta}_j}$ veces el riesgo al evento de las unidades que no pertenecen al grupo, manteniendo fijas las demás covariables.

6.3.2.1. Intervalos de confianza y verificación de hipótesis sobre los β's

Los errores estándar de los estimadores y la distribución asintótica de los $\hat{\beta}_j$, $j = 1, 2, \ldots, p$, pueden emplearse para la construcción de intervalos o regiones de confianza y para verificar hipótesis sobre los respectivos parámetros.

Así, un intervalo de confianza del $(1-\alpha)100\,\%$ para β_j es

$$\hat{\beta}_j \mp Z_{(1-\alpha/2)}ee(\hat{\beta}_j),$$

donde $Z_{1-\alpha/2}$ es el percentil $1-\alpha/2$ de una normal estándar y $ee(\hat{\beta}_j)$ es el error estándar de $\hat{\beta}_j$.

Si un intervalo $(\hat{\beta}_j - Z_{(1-\alpha/2)}ee(\hat{\beta}_j),\ \hat{\beta}_j + Z_{(1-\alpha/2)}ee(\hat{\beta}_j))$ no contiene al cero, esto es una evidencia de que el respectivo parámetro β_j es diferente de cero, y en consecuencia, la respectiva covariable tiene un efecto sobre el riesgo al evento.

La verificación de la hipótesis $H_0:\ \beta_j = 0$ puede desarrollarse mediante la estadística $z_0\hat{\beta}_j/ee(\hat{\beta}_j)$, cuya región crítica o valor p puede determinarse con la distribución normal estándar; es decir, $p = P(|\ Z\ |> z_0)$. Para un valor establecido de α, se rechaza H_0 si $p < \alpha$, en caso contrario, no se rechaza.

Para verificar la hipótesis de que "ninguna covariable tiene efecto sobre el riesgo al evento", es decir, $H_0 : \beta = (\beta_1, \ldots, \beta_p)' = 0$, se emplea la estadística de Wald:

$$W = (\hat{\beta} - \beta_0)' \left(\mathcal{IO}(\hat{\beta})\right) (\hat{\beta} - \beta_0),$$

la cual tiene una distribución ji-cuadrado con p-grados de libertad. Para la hipótesis de que "ninguna covariable tiene efecto sobre el riesgo al evento", $\beta_0 = 0$. Si se rechaza H_0, entonces se concluye que al menos una de las p covariables tiene un efecto sobre el riesgo al evento. El procedimiento adecuado debería ser, primero, verificar la hipótesis global de nulidad del vector de parámetros β, y luego, si se rechaza esta hipótesis de nulidad, verificar cuál o cuáles covariables muestran un efecto significativo sobre el riesgo al evento. En la sección 6.4 se desarrollan los procedimientos estadísticos más apropiados para verificar la necesidad o no de incluir algunas covariables en el modelo.

6.3.2.2. Intervalos de confianza para la razón de riesgos

La razón de riesgos al evento en un tiempo t para dos unidades que difieren en las covariables únicamente en la pertenencia a un grupo es igual a e^{β_j}, donde la covariable $x_j = 1$, si la unidad pertenece al grupo, y $x_j = 0$, si no pertenece al grupo. Se escribe $\hat{\psi} = e^{\hat{\beta}_j}$, por la propiedad de invarianza de los estimadores máximo verosímiles. El error estándar de $\hat{\psi}$ puede ser obtenido mediante el Método Delta (2.4.3) así:

$$var(\hat{\psi}) \approx \hat{\psi}^2 var(\hat{\beta}_j) = \left(e^{\hat{\beta}_j}\right)^2 var(\hat{\beta}_j). \tag{6.3.11}$$

Luego, el error estándar de $\hat{\psi}$ es aproximado por

$$ee(\hat{\psi}) \approx \hat{\psi}\, ee(\hat{\beta}_j). \tag{6.3.12}$$

Un intervalo del $(1-\alpha)100\,\%$ de confianza para ψ es dado por

$$\widehat{\psi} \mp \mathcal{Z}_{(1-\alpha/2)} ee(\widehat{\psi}). \tag{6.3.13}$$

Como se trata de una razón o cociente de riesgos, se debe observar si el intervalo de confianza contiene al uno. Si lo contiene significa que el riesgo al evento es igual para las unidades del grupo como para las que no pertenecen al grupo. Si no contienen al uno, hay dos posibilidades. Una es que ambos extremos del intervalo sean menores que uno, en cuyo caso se puede afirmar que tienen más riesgo al evento las unidades que no pertenecen al grupo y en

este caso, al grupo se le denomina "protector" al evento. La otra posibilidad es que ambos extremos del intervalo sean mayores que uno, entonces, el grupo "protector" es el complemento del grupo.

El grupo puede estar conformado por las unidades que reciben un tratamiento (pacientes que reciben un medicamento experimental), mientras que su complemento (pacientes que reciben un placebo) lo conforman las unidades que no reciben tal tratamiento.

6.3.3. Estimación de funciones relacionadas con $h_0(t)$

Además de la estimación de los parámetros β, se requiere estimar funciones relacionadas con $h_0(t)$ para determinar completamente el modelo de Cox. Una de estas funciones, relacionada con la función de riesgo base, es la tasa de riesgo al evento acumulada base:

$$H_0(t) = \int_0^t h_0(u)du. \tag{6.3.14}$$

La correspondiente función de sobrevida base es:

$$S_0(t) = \exp\{-H_0(t)\}. \tag{6.3.15}$$

Estas funciones son importantes como herramientas para evaluar la adecuación de un modelo, tal como los residuos de Cox-Snell y otros que se presentan más adelante.

La función de sobrevida

$$S(t|\mathbf{x}) = [S_0(t)]^{\exp\{\mathbf{x}'\beta\}}$$

es útil cuando se quiere analizar en términos de percentiles asociados con grupos de unidades.

Como no se ha supuesto un modelo probabilístico para la variable *tiempo al evento*, T, no se puede acudir a la función de verosimilitud para estimar $h_0(t)$. Además, el método de verosimilitud parcial elimina esta función, de manera que se debe acudir a procedimientos no paramétricos para la estimación de $h_0(t)$. Una alternativa simple propuesta por Breslow [5] para estimar $H_0(t)$ es una función escalonada con saltos en los tiempos diferentes en los que ocurre el evento; esta se expresa como:

$$\widehat{H}_0(t) = \sum_{j:t_j<t} \frac{e_j}{\sum_{l\in R(t_j)} \exp\left\{\mathbf{x}'\hat{\beta}\right\}}, \tag{6.3.16}$$

donde e_j es el número de eventos en el tiempo t_j. Desde aquí, se pueden estimar las funciones de sobrevida $S_0(t)$ y $S(t)$, respectivamente, mediante

$$\widehat{S}_0(t) = \exp\left\{-\widehat{H}_0(t)\right\}, \tag{6.3.17}$$

y

$$\widehat{S}(t) = \widehat{S}_0(t)^{\exp\{\mathbf{x}'\hat{\beta}\}}. \tag{6.3.18}$$

En ausencia de covariables, la expresión (6.3.16) se reduce a

$$\widehat{H}_0(t) = \sum_{j:t_j<t} \left(\frac{e_j}{n_j}\right), \tag{6.3.19}$$

que corresponde al estimador de Nelson-Aalen para $H(t)$. Este estimador también se conoce con el nombre de estimador de Nelson-Aalen-Breslow.

Ejemplo 6.3.1. Los datos de la tabla 7.9 son los tiempos en meses de sobrevida de mujeres que han recibido una mastectomía simple o radical (extirpación parcial o total de una glándula mamaria) para tratar un tumor. Los tiempos de sobrevida de cada mujer se clasifican de acuerdo a si el tumor es identificado (de acuerdo con un examen o prueba) como positivo o negativo. Los datos censurados a derecha se rotulan con el signo más ("+").

La variable que indica el resultado del diagnóstico es un factor con dos niveles (positivo y negativo). Este factor se puede incluir como covariable en el modelo de Cox empleando una variable indicadora X, donde $X = 0$ significa un diagnóstico con resultado negativo y $X = 1$, un resultado con diagnóstico positivo. Bajo el modelo de riesgos proporcionales de Cox, el riesgo al evento muerte en el tiempo t para la i-ésima mujer, cuyo valor de covariable es x_i, es

$$h(t|x_i) = h_0(t)e^{\beta x_i},$$

con $x_i = 0, 1$. La función $h_0(t)$ es la función de riesgo cuando un tumor fue diagnosticado negativamente. En el grupo de mujeres que fueron diagnosticadas negativamente, hay dos con tiempos de sobrevida empatados en 26 meses. La aproximación a la función de verosimilitud de Breslow (6.3.10) se emplea para este caso.

Tabla 6.3. Tiempo de sobrevida en mujeres con cáncer de mama

Resultado del examen diagnóstico		
Negativo	Positivo	Positivo
23	5	68
47	8	71
69	10	76^+
70^+	13	105^+
71^+	18	107^+
100^+	24	109^+
101^+	26	113
148	26	116^+
181	31	118
198^+	35	143
208^+	40	154^+
212^+	41	162^+
224^+	48	188^+
	50	212^+
	59	217^+
	61	225^+

Tomado de Collet [10, p. 7].

La siguiente es la sintaxis con la cual se procesan los datos en R para ajustar el modelo de Cox. Se emplea la función "coxph" del paquete *survival*; los empates se consideran con la metodología de Breslow.

```
t<-c(23,47,69,70,71,100,101,148,181,198,208,212,224,5,8,
+ 10,13,18,24,26,26,31,35,40,41,48,50,59,61,68,71,
+ 76,105,107,109,113,116,118,143,154,162,188,212,217,225)
cen<-c(1,1,1,0,0,0,0,1,1,0,0,0,0,rep(1,18),rep(0,4),1,0,1,1,
+rep(0,6))
grupo<-c(rep(0,13),rep(1,32))
datos<-cbind(t,cen,grupo)
datos<-as.data.frame(datos)
require(survival)
cox1<-coxph(Surv(t,cen)~factor(grupo),data=datos,x =T,
+method="breslow")
```

El estimador máximo verosímil de β es $\hat{\beta} = 0.908$, con un error estándar de $ee(\hat{\beta}) = 0.501$. La razón de riesgos es $e^{\hat{\beta}} = e^{0.908} = 2.4794$.

Como es mayor que uno, se concluye que las mujeres con pronóstico positivo tienen un riesgo de muerte de 2.5 veces el que tienen las mujeres con pronóstico negativo.

El error estándar de la razón de riesgos puede encontrarse mediante el Método Delta (6.3.11), usando el error estándar de $\hat{\beta}$. La estimación del riesgo relativo es $\hat{\psi} = e^{\hat{\beta}} = 2.4794$ y el error estándar de $\hat{\beta}$ es 0.501. Por el Método Delta el error estándar de $\hat{\psi}$ es

$$ee(\hat{\psi}) = 2.4794 \times 0.501 = 1.2422.$$

Un intervalo del 95 % de confianza para β es: $\hat{\beta} \mp 1.96 ee(\hat{\beta})$, lo que corresponde al intervalo $(-0.07396; 1.88996)$. Luego, un intervalo de confianza para $\psi = e^{\beta}$ se obtiene por la exponenciación del intervalo anterior, este es igual a $(0.9289, 6.618)$. Note que este intervalo "escasamente" contiene al uno, lo cual advierte que los dos grupos de mujeres tienen diferente riesgo de muerte.

Ejemplo 6.3.2. Los datos de mielanoma múltiple (cáncer en médula ósea) registrados en 48 pacientes que padecen esta enfermedad se consignan en la tabla 6.4. Los datos están conformados por los valores de siete variables que fueron medidas sobre cada uno de los 48 pacientes. La notación de las variables es la siguiente:

- Edad: edad del paciente en años.
- Sexo: sexo del paciente (0 es hombre y 1 es mujer).
- Bun: nitrógeno ureico en sangre (por su sigla en inglés).
- Ca: calcio en sangre.
- Hb: hemoglobina en sangre.
- Pcells: porcentaje de células en plasma.
- Protein: proteína Bence-Jones (0 es ausencia y 1 es presencia).

El modelo de Cox en forma explícita para el paciente i-ésimo, en las covariables $\mathbf{x}_i = (Edad, Sexo, Bun, Ca, Hb, Pcells, Protein)_i'$, se expresa como:

$$h(t|\mathbf{x_i}) = h_0(t) \exp\{\beta_1 edad_i + \beta_2 Sexo_i + \beta_3 Bun_i + \beta_4 Ca_i + \beta_5 Hb_i + \beta_6 Pcells_i + \beta_7 Protein_i\}. \quad (6.3.20)$$

El valor de la función de riesgo base, $h_0(t)$, corresponde a la función de riesgo para un paciente a quien no se le registran covariables. No es correcto decir que corresponde al riesgo de un paciente para quien estas siete

Tabla 6.4. Tiempo de sobrevida de pacientes que padecen melanoma

Paciente	Tpo	cen	Edad	Sexo	Bun	Ca	Hb	Pcells	Protein
1	13	1	66	1	25	10	14.6	18	1
2	52	0	66	1	13	11	12.0	100	0
3	6	1	53	2	15	13	11.4	33	1
4	40	1	69	1	10	10	10.2	30	1
5	10	1	65	1	20	10	13.2	66	0
6	7	0	57	2	12	8	9.9	45	0
7	66	1	52	1	21	10	12.8	11	1
8	10	0	60	1	41	9	14.0	70	1
9	10	1	70	1	37	12	7.5	47	0
10	14	1	70	1	40	11	10.6	27	0
11	16	1	68	1	39	10	11.2	41	0
12	4	1	50	2	172	9	10.1	46	1
13	65	1	59	1	28	9	6.6	66	0
14	5	1	60	1	13	10	9.7	25	0
15	11	0	66	2	25	9	8.8	23	0
16	10	1	51	2	12	9	9.6	80	0
17	15	0	55	1	14	9	13.0	8	0
18	5	1	67	2	26	8	10.4	49	0
19	76	0	60	1	12	12	14.0	9	0
20	56	0	66	1	18	11	12.5	90	0
21	88	1	63	1	21	9	14.0	42	1
22	24	1	67	1	10	10	12.4	44	0
23	51	1	60	2	10	10	10.1	45	1
24	4	1	74	1	48	9	6.5	54	0
25	40	0	72	1	57	9	12.8	28	1
26	8	1	55	1	53	12	8.2	55	0
27	18	1	51	1	12	15	14.4	100	0
28	5	1	70	2	130	8	10.2	23	0
29	16	1	53	1	17	9	10.0	28	0
30	50	1	74	1	37	13	7.7	11	1
31	40	1	70	2	14	9	5.0	22	0
32	1	1	67	1	165	10	9.4	90	0
33	36	1	63	1	40	9	11.0	16	1
34	5	1	77	1	23	8	9.0	29	0
35	10	1	61	1	13	10	14.0	19	0
36	91	1	58	2	27	11	11.0	26	1
37	18	0	69	2	21	10	10.8	33	0
38	1	1	57	1	20	9	5.1	100	1
39	18	0	59	2	21	10	13.0	100	0
40	6	1	61	2	11	10	5.1	100	0
41	1	1	75	1	56	12	11.3	18	0
42	23	1	56	2	20	9	14.6	3	0
43	15	1	62	2	21	10	8.8	5	0
44	18	1	60	2	18	9	7.5	85	1
45	12	0	71	2	46	9	4.9	62	0
46	12	1	60	2	6	10	5.5	25	0
47	17	1	65	2	28	8	7.5	8	0
48	3	0	59	1	90	10	10.2	6	1

Tomado de Collet [10, p. 9].

variables toman el valor cero, pues en el caso de covariables tipo factor o categóricas habría ambigüedad.

A continuación, se muestra la salida del procesamiento con R mediante la función "coxph". La salida contiene la estimación de los siete parámetros, $\hat{\beta}_j$, los $e^{\hat{\beta}_j}$, los errores estándar de las estimaciones, $ee(\hat{\beta}_j)$, el valor de la estadística normal, z, y el valor p asociado con z y los intervalos de 95 % confianza para e^{β_j}.

```
coxph(formula = Surv(collet$Tpo, collet$cen) ~ collet$edad +
factor(sexo) + Bun + Ca + Hb + Pcells + factor(Protein),
data = collet, x = T, method = "breslow")

n= 48, number of events= 36

                    coef     exp(coef)  se(coef)      z Pr(>|z|)
collet$edad      -0.019358  0.980828  0.027924 -0.693 0.488159
factor(sexo)2    -0.250899  0.778101  0.402286 -0.624 0.532836
Bun             0.020826  1.021044  0.005929  3.513 0.000443 ***
Ca                0.013125  1.013211  0.132442  0.099 0.921061
Hb              -0.135241  0.873506  0.068891 -1.963 0.049635 *
Pcells           -0.001594  0.998407  0.006577 -0.242 0.808533
factor(Protein)1 -0.640438  0.527061  0.426687 -1.501 0.133367
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

                    exp(coef) exp(-coef) lower .95 upper .95
collet$edad         0.9808     1.0195    0.9286    1.0360
factor(sexo)2       0.7781     1.2852    0.3537    1.7118
Bun                 1.0210     0.9794    1.0092    1.0330
Ca                  1.0132     0.9870    0.7816    1.3135
Hb                  0.8735     1.1448    0.7632    0.9998
Pcells              0.9984     1.0016    0.9856    1.0114
factor(Protein)1    0.5271     1.8973    0.2284    1.2163
```

De acuerdo con los intervalos de confianza (y los valores p) para cada uno de los siete parámetros, de la salida de R anterior se puede afirmar, con una confiabilidad de 95 %, que *Bun* y *Hb* son significativamente diferentes de cero. Con este resultado, sobre los datos disponibles, se puede advertir que la función de riesgo no depende de estas siete variables.

No obstante, no es correcto afirmar, a partir de estos resultados, que *Bun* y *Hb* son variables relevantes, dado que las estimaciones de los coeficientes de las siete variables explicativas en el modelo no son independientes.

Para determinar sobre cuáles covariables depende la función de riesgo, se requeriría de diferentes modelos. Para el caso de estas siete covariables se debería "ensayar" con $2^7 = 128$ modelos diferentes (sin considerar posibles interacciones). Estos van desde un modelo sin covariables (modelo nulo), siete modelos con una sola covariable, hasta un modelo con las siete covariables (modelo saturado) como el referenciado en la expresión (6.3.20). En la siguiente sección se resumen algunas estrategias para comparar el ajuste de diferentes modelos.

6.4. Adecuación del modelo de Cox

Como la mayoría de los modelos estadísticos, el modelo de Cox requiere el uso de herramientas exploratorias y confirmativas para dar cuenta de su adecuación o ajuste a un conjunto de datos inscritos en un marco conceptual específico. En esta etapa se debe verificar que los supuestos con los cuales se estructura y estima el modelo se cumplen en los datos. En particular, un supuesto básico es la proporcionalidad en los riesgos. El no cumplimiento de este supuesto puede conllevar a obtener estimaciones de los coeficientes del modelo con problemas de sesgo, varianza, consistencia, entre otros.

En esta sección se muestran algunos métodos para evaluar la adecuación de este modelo, los cuales se encuentran en la mayoría de la literatura de modelos de regresión. Además, se presentan algunos métodos propios para los modelos de regresión de datos de tiempo hasta el evento.

6.4.1. Métodos gráficos y residuos

Para verificar el supuesto de riesgos proporcionales en el modelo de Cox, un gráfico es usado para explorar el supuesto de proporcionalidad. Se dividen los datos en un número determinado de estratos, usualmente de acuerdo con alguna covariable categórica o con la categorización de una variable numérica. Por ejemplo, dividir los datos en dos estratos de acuerdo con la covariable *sexo*. A continuación, se debe estimar la función de riesgo base acumulada $H_0(t)$ para cada estrato mediante la expresión (6.3.16). Si el supuesto es válido, las curvas del logaritmo $\widehat{H}_0(t)$ frente a t o $\ln(t)$ deben presentar diferencias aproximadamente constantes en el tiempo. Las curvas no paralelas significan que el supuesto de riesgos proporcionales no se cum-

ple, como pasa, por ejemplo, en la situación mostrada en el panel derecho de la figura 6.1. Una ventaja de esta técnica gráfica es advertir cuál es la covariable que está asociada con el no cumplimiento de la proporcionalidad, en este caso, una variable categórica o categorizada.

Una herramienta alternativa para explorar y verificar la no existencia de riesgos proporcionales en los datos incluidos en el modelo de Cox es el análisis de los *residuos de Schoenfeld*, propuestos en Schoenfeld [63]. Estos residuos se definen para cada unidad y cada covariable, y se basan en la primera derivada de la función de log-verosimilitud.

Hay varios métodos gráficos para diagnosticar la bondad de ajuste de un modelo de riesgos proporcionales. Los métodos gráficos se basan en los residuos y se emplean como herramientas de diagnóstico. En los modelos de regresión clásicos, los residuos se refieren a la diferencia entre un valor observado y un valor pronosticado por el modelo de regresión ajustado en la variable respuesta. En observaciones censuradas únicamente cuando se emplea la verosimilitud parcial en modelos de riesgos proporcionales, el concepto usual de residuo no es aplicable. Además de los residuos de Cox-Snell presentados en la sección (4.3.0.8), se definen tres residuos de interés en el diagnóstico del modelo de Cox; estos son los de Schoenfeld, Martingala y Deviance.

6.4.1.1. Residuos de Schoenfeld

Si la unidad $i = 1, 2 \ldots, n$ es representada por el vector que contiene p covariables $\mathbf{x}_i = (x_{1i}, x_{2i} \ldots, x_{pi})'$ y sobre esta unidad ocurre el evento, se tiene un vector de *residuos Schoenfeld*, $(r_{S1i}, r_{S2i}, \ldots, r_{Spi})$, donde cada componente de este vector es definido por

$$r_{Sqi} = \delta_i \left[x_{qi} - \frac{\sum_{j \in R(t_i)} x_{qj} \exp\{\mathbf{x}_j' \hat{\beta}\}}{\sum_{j \in R(t_i)} \exp\{\mathbf{x}_j' \hat{\beta}\}} \right], \; q = 1, 2, \ldots, p, \; i = 1, 2, \ldots, n, \tag{6.4.1}$$

con $\delta_i = 1$ si se observa el tiempo al evento t_i y $\delta_i = 0$ si hay censura; $\hat{\beta}$ es el estimador de máxima verosimilitud parcial de β. Los residuos de Schoenfeld se definen únicamente para los tiempos no censurados y las observaciones censuradas se consideran como datos faltantes.

Si hay $k \leq n$ tiempos diferentes, entonces los residuos conforman una matriz de tamaño $k \times p$, donde cada fila contiene los residuos de cada unidad calculados en las p-covariables incluidas en el modelo.

Como en la mayoría de los residuos, se tiene que $\sum_i r_i = 0$ y asintóticamente estos residuos tienen media cero. También se puede demostrar que estos residuales no están correlacionados.

Se definen los *residuos de Schoenfeld escalados* como

$$r_{Si}^{*} = e \times \mathcal{I}O(\hat{\beta})^{-1} \times r_{Si}, \tag{6.4.2}$$

donde e es el número de eventos entre las n unidades y $\mathcal{I}O(\hat{\beta})$ es la matriz de información observada.

Los gráficos de los residuos de Schoenfeld frente al tiempo o a una covariable pueden usarse para inspeccionar la adecuación del modelo de riesgos proporcionales. La presencia de algún patrón en la gráfica advierte la no proporcionalidad de los riesgos al evento, mientras que los distanciamientos extremos del grupo principal de residuos señalan la presencia de datos atípicos o posibles problemas de estabilidad del modelo.

6.4.1.2. Residuos de Martingala

Los *residuos de Martingala* se definen por

$$r_{Mi} = \delta_i - r_{CSi}, \ i = 1, 2, \ldots, n, \tag{6.4.3}$$

con $\delta_i = 1$ si el tiempo al evento, t_i, es observado (no censurado), y 0 si es censurado; r_{CSi} son los residuos de Cox-Snell. Estos residuos, que son una modificación de los residuos de Cox-Snell, se consideran como una estimación del número de eventos observados en exceso en los datos, pero no pronosticados por el modelo. Se emplean para examinar la forma funcional (lineal, cuadrática, etc.) para una covariable del modelo de regresión. Así, en una gráfica de dispersión de los puntos (x_{ij}, r_{Mi}) para $i = 1, 2, \ldots, n$ y x_j una covariable continua, la nube de puntos insinúa la transformación que debe hacerse sobre esta variable.

La distribución de los residuos de Martingala tiene una distribución sesgada con media cero.

6.4.1.3. Residuos de Deviance

En la mayoría de la literatura estadística se ha mantenido la palabra *deviance* sin traducción particular. Se definen los *residuos de Deviance* como

$$r_{Di} = sign(r_{Mi})\sqrt{-2[r_{Mi} + \delta_i \ln(\delta_i - r_{Mi})]}, \ i = 1, 2, \ldots, n, \tag{6.4.4}$$

donde $sign(\cdot)$ es la función signo, la cual toma el valor 1 si el argumento es positivo, 0 si el argumento es cero y -1 si el argumento es negativo; r_{Mi} es el residuo de Martingala.

Estos residuos son una alternativa a los Martingala, pues son más simétricos en torno a cero y, en general, facilitan la detección de observaciones

atípicas (*outliers*). Para un modelo que se ajuste adecuadamente a los datos, estos residuos deben presentar una nube de puntos aleatoria en torno a cero. Una gráfica de los residuos de Martingala y Deviance frente al tiempo, además de dar cuenta de la adecuación del modelo ajustado, permite la detección de datos atípicos.

6.4.2. Comparación de modelos alternativos

En el análisis de regresión para los datos de *tiempo hasta un evento* se modela la dependencia de la función de riesgo sobre un conjunto de covariables o variables explicativas. En este sentido, se ajustan los modelos de riesgos proporcionales con un componente lineal que contiene las covariables; se trata de comparar los diferentes modelos. Como se ha señalado, si se dispone de p covariables, entonces se tienen 2^p posibles modelos, que van desde el modelo sin covariables (modelo nulo) hasta el modelo con todas las p covariables (modelo saturado).

Un primer modelo está *anidado* en un segundo modelo si todos los términos del primero están incluidos en el segundo. Así, para el modelo de Cox, $h(t \mid x) = h_0(t)e^{\mathbf{x}'\beta}$, donde el modelo nulo es $h_0(t)$. Para el caso de un factor o una variable dicotómica, $x = 0, 1$ (no tratamiento y tratamiento, respectivamente), los modelos son $h_0(t)$ y $h_0(t)e^{\beta}$, respectivamente.

En general, suponga que se tiene el Modelo 1 y el Modelo 2, donde el primero contiene un subconjunto de covariables del segundo. Estos son: Modelo 1,

$$h(t \mid \mathbf{x}_1) = h_0(t) \exp\left\{\beta_1 X_1 + \cdots + \beta_p X_p\right\},$$

y Modelo 2,

$$h(t \mid \mathbf{x}) = h_0(t) \exp\left\{\beta_1 X_1 + \cdots + \beta_p X_p + \beta_{p+1} X_{p+1} + \cdots + \beta_{p+q} X_{p+q}\right\}.$$

Se esperaría que el Modelo 2 tenga un mejor ajuste que el Modelo 1. El problema es determinar si las q variables adicionales o "extras" aumentan significativamente el poder de predicción y explicación del Modelo 2; en caso contrario, se debería asumir el Modelo 1.

El efecto de una covariable no puede ser estudiado independientemente de las demás covariables. Así, el efecto de una covariable depende de las otras covariables incluidas en el modelo. Por ejemplo, en el Modelo 1, el efecto de cualquiera de las p variables sobre la función de riesgo depende de las otras $p - 1$ covariables que han sido ajustadas al modelo; por lo tanto, se dice que el efecto de la variable X_p está ajustado por la restantes $p - 1$ covariables. De manera similar, cuando q covariables $X_{p+1}, X_{p+2}, \ldots, X_{p+q}$ se

adicionan al modelo, se dice que el efecto de estas covariables sobre la función de riesgo está ajustado por las p covariables que ya han sido ajustadas, $X_1, X_2, \ldots, X_p$.

Para comparar los modelos alternativos ajustados a los mismos datos observados, se requiere de una estadística que mida la extensión a la cual los datos son ajustados por un modelo particular. Como la función de verosimilitud resume la información que los datos contienen sobre los parámetros desconocidos en un modelo dado, una estadística apropiada es el valor de dicha función de verosimilitud cuando los parámetros son reemplazados por sus estimadores de máxima verosimilitud. Para un conjunto de datos particular, entre más grande el valor de la verosimilitud maximizada, mejor es el ajuste entre el modelo y los datos.

La estadística que mide el acuerdo (ajuste) entre el modelo y los datos es $-2\ln\widehat{L}$. Como se describe en la sección 6.3, $\widehat{L}$ es el producto de probabilidades condicionales y, por lo tanto, esta estadística toma valores entre cero y uno; en consecuencia, $-2\ln\widehat{L}$ es siempre positivo. Esta expresión se obtiene en la aproximación límite (asintótica) de esta estadística como se muestra en Hogg, McKean y Craig [32, p. 353]. El modelo que tenga *el menor valor* de $-2\ln\widehat{L}$ se considera como el mejor modelo. Esta estadística es útil unicamente cuando se hacen comparaciones entre modelos ajustados para los mismos datos; es decir, si se amplia o se reduce la base de datos, esta estadística puede cambiar para el mismo modelo.

Considere el Modelo 1 y el Modelo 2 definidos anteriormente, y sean $\widehat{L}(1)$ y $\widehat{L}(2)$ el valor del logaritmo de la verosimilitud maximizada para los modelos 1 y 2, respectivamente. Los dos modelos se pueden comparar sobre la base de las diferencias entre $-2\ln\widehat{L}(1)$ y $-2\ln\widehat{L}(2)$, lo que permite concluir si las q variables en el Modelo 2, que son adicionales a las que ya están en el Modelo 1, aportan al ajuste del modelo a los datos. De otra forma, la diferencia de los logaritmos de las verosimilitudes equivale a

$$-2\ln\left(\frac{\widehat{L}(1)}{\widehat{L}(2)}\right)\underset{\rightarrow}{D}\ \chi_{(q)}.$$

Esta es la estadística con la cual se verifica la hipótesis $H_0: \beta_{p+1} = \cdots = \beta_{p+q} = 0$. Los valores grandes de esta estadística advierten sobre la necesidad de incluir parámetros (covariables) adicionales a los ya contenidos en el Modelo 1.

6.4.2.1. Selección dentro del modelo

En la construcción de cualquier modelo, y en particular del modelo de Cox, uno de los primeros pasos es decidir el conjunto de covariables que harán parte de su componente lineal. El conjunto de covariables puede estar conformado por factores y variables que están medidos o registrados sobre las unidades. Además, algunas interacciones entre las covariables pueden requerirse en el componente lineal del modelo de Cox; es decir, en

$$\psi(\mathbf{x}) = exp\{\mathbf{x}'\beta\} = \exp\left\{\beta_1 X_1 + \cdots + \beta_p X_p\right\}.$$

Una vez que se ha decidido sobre el conjunto de covariables a incluir en el modelo, el paso siguiente es seleccionar la combinación de las covariables que deben usarse en el modelo de riesgo. De manera que, de acuerdo con cada combinación de las covariables, se obtiene un modelo. Entre estos modelos habrán unos "mejores" que otros, el problema es decidir cuál es el mejor modelo.[1]

En modelamiento, se sigue el *principio de jerarquía*, lo que significa que si un modelo tiene un término de interacción entre covariables, entonces todas las covariables que conforman la interacción deben estar incluidas en el modelo. Esto implica que los términos de interacción no deben ser ajustados a menos que las covariables que lo componen (efectos principales) estén presentes en el modelo.

Una vez que se ha estructurado un modelo, manteniendo el principio de jerarquía, se debe decidir sobre los términos del predictor lineal (factores, variables e interacciones) que deben incluirse en la función de riesgo.

6.4.2.2. Métodos de selección de variables

Se inicia con un modelo en el cual todas las covariables se consideran e incluyen. El objetivo es identificar los subconjuntos de covariables apropiadas para el modelo de riesgo proporcional. Para un número de covariables, interacciones e incluso términos no lineales que no sea demasiado grande, puede ser posible incluirlos en el modelo bajo el principio de jerarquía. Sin embargo, esto también depende de la cantidad de observaciones disponibles, es decir, del tamaño de muestra (esto se trata en el capítulo 11). Los modelos alternativos anidados se pueden comparar mediante la estadística $-2\ln\widehat{L}$, a través de la inclusión (o supresión) de términos en uno de los modelos.

[1] "No existe el mejor modelo, hay unos menos malos que otros": George Box.

Una comparación entre un número posible de modelos, los cuales no necesariamente son anidados o encajados, se puede hacer con base en la estadística

$$AIC = -\ln \widehat{L} + \alpha q, \tag{6.4.5}$$

donde q es el número de parámetros desconocidos en el modelo y α es una constante predeterminada, generalmente con valores entre 2 y 6. Un valor de α igual 3 es similar al 5 % empleado en la significancia de la verificación de hipótesis. La estadística AIC, por sus siglas en inglés (Akaike's Information Criterion), se conoce como el criterio de información de Akaike. Los modelos asociados con valores relativamente pequeños de esta estadística califican al modelo como el "mejor". Una característica, no explícita, de esta estadística es que si la única diferencia entre dos modelos es que uno incluye covariables no necesarias, el valor de AIC para ambos modelos puede no ser diferente. El valor de AIC tenderá a incrementarse cuando términos innecesarios se adicionen al modelo.

Cuando se incluyen covariables en un modelo, es necesario haber explorado previamente la posible dependencia (colinealidad) entre estas, para no poner covariables redundantes en el modelo que le resten poder de predicción y explicación. Si el número de covariables es grande, puede resultar computacionalmente costoso el ajuste de todos los posibles modelos, pues hay 2^p posibles combinaciones de términos (un modelo por cada combinación). Así, con $p = 6$, se tendrían $2^6 = 128$ modelos, mientras que con $p = 10$ se tendrían $2^{10} = 1024$ modelos, y así sucesivamente. De esta manera, se requiere de técnicas de computación para la selección de variables, las cuales se encuentran disponibles en la mayoría de los paquetes estadísticos. Estas herramientas de cómputo se basan en la *selección hacia adelante*, *eliminación hacia atrás* o una combinación de estas dos, como la *búsqueda paso a paso*. Estas se conocen por sus palabras en inglés *forward, backward* y *stepwise*, respectivamente.

Para la *selección hacia adelante*, las covariables se adicionan al modelo una a la vez. En cada etapa del proceso, la covariable adicionada al modelo es aquella que, al ser incluida, suministre la mayor disminución en el valor de la estadística $-2 \ln \widehat{L}$. El proceso termina cuando una covariable candidata a incluirse en el modelo no reduzca el valor de $-2 \ln \widehat{L}$ en una cantidad previamente especificada. Este valor se conoce como la "regla de parada". La mayoría de los paquetes estadísticos mencionados tiene la opción de declarar este valor en la programación del cálculo computacional; en caso de no explicitarlo, el programa de cómputo lo hace por defecto.

En la *eliminación hacia atrás*, un modelo que contiene el mayor número de covariables es ajustado. Luego, las covariables se van excluyendo de una en

una. En cada etapa, la variable excluida es aquella que incremente el valor de $-2\ln\widehat{L}$ en la menor cantidad por su exclusión. El proceso termina cuando la siguiente candidata a excluir incrementa el valor de $-2\ln\widehat{L}$ en una cantidad superior a un valor preestablecido.

El procedimiento de *búsqueda o selección paso a paso* opera de la misma forma que la selección hacia adelante. No obstante, una variable que en una etapa anterior ha sido incluida en el modelo puede ser considerada para su exclusión en etapas posteriores. El proceso de parada tiene también un valor establecido con anterioridad.

Ejemplo 6.4.1. Los datos corresponden a 38 pacientes a los que se les registró el tiempo de sobrevida en meses. Los tiempos de vida de pacientes que murieron por otras causas, o a quienes se les perdió el seguimiento durante el estudio, se consideran como censurados; una variable indicadora establece si el tiempo es observado ($\delta_i = 1$) o censurado a derecha ($\delta_i = 0$). La variable tratamiento (*Trat.*) toma el valor 2 cuando la persona es tratada con DES (dietilestilbestrol, que es un estrógeno sintetizado) y 1 si la persona toma un placebo. El objetivo del estudio es determinar si los pacientes tratados con DES tienen un tiempo de sobrevida mayor que los tratados con el placebo. Como los datos corresponden a un experimento aleatorizado, se puede esperar que la distribución de los factores de pronóstico, tales como la edad del paciente, el nivel de hemoglobina en suero sanguíneo (*Hemog.*), el tamaño del tumor (*TTumor*) y el índice de Gleason (*IGleason*), será similar en los pacientes en cada uno de los dos grupos de tratamiento [10, p. 10]. En consecuencia, una de las primeras tareas es determinar cuál o cuáles covariables se relacionan con el tiempo de sobrevida.

Todas las posibles combinaciones (subconjuntos) de estas variables se ajustan en un modelo de Cox, y se calculan los valores del logaritmo de la razón de verosimilitudes, $-2\ln\widehat{L}$, así como los valores del criterio de información de Akaike, AIC (con $\alpha = 2$).

```
library(MASS)
collet14<-read.table("Tabla1_4_Collet.txt", header=T)
attach(collet14)
collet14<-as.data.frame(collet14)

require(survival)
cox<-coxph(Surv(collet14$Tpo,collet14$Cen)~SHeam+Ttumor,
+data=collet14, x=T,method="breslow")
RV<-anova(cox)
RV
Reg_paso<-stepAIC(cox)
```

Tabla 6.5. Tiempo de sobrevida en pacientes con cáncer de próstata

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

La tabla 6.6 contiene estos resultados. De acuerdo con esta tabla, los modelos más significativos que incluyen una sola covariable son los asociados con el tamaño del tumor (Ttumor) y el índice de Gleason (IGleason). A

Tabla 6.6. Modelos ajustados para los datos de cáncer de próstata

Modelo Núm.	Modelo	$-2\ln\widehat{L}$	AIC
M1	Nulo	36.348	36.349
M2	Edad	36.270	38.269
M3	Hemog	36.196	38.196
M4	Ttumor	29.042	31.042
M5	IGleason	29.126	31.126
M6	Edad+Hemog	36.150	40.151
M7	Edad+Ttumor	28.854	32.854
M8	Edad+IGleason	28.760	32.760
M9	Hemog+Ttumor	29.019	33.019
M10	Hemog+Gleason	27.982	31.982
M11	Ttumor+IGleason	23.534	27.534
M12	Edad+Hemog+TTumor	28.852	34.852
M13	Edad+Hemog+IGleason	27.894	33.894
M14	Edad+Ttumor+IGleason	23.270	29.270
M15	Hemog+Ttumor+IGleason	23.508	29.508
M16	Edad+Hemog+Ttumor+IGleason	23.230	31.231

partir del cambio en el valor de $-2\ln\widehat{L}$ al omitir cada una de estas variables en un modelo que contiene a las dos variables, se concluye que ambas son necesarias en un modelo de riesgos proporcionales. El valor de $-2\ln\widehat{L}$ es únicamente reducido por una pequeña cantidad cuando la covariable *Edad* y *Hemog* se adicionan al modelo que contiene las covariables *Ttumor* e *IGleason*. Se concluye que únicamente las covariables *Ttumor* e *IGleason* son importantes para el pronóstico.

Respecto a la estadística AIC, la tabla 6.6 muestra que el modelo con las covariables Ttumor e IGleason tiene asociado el menor valor de esta estadística (27.534), lo cual confirma que, entre los 16 modelos propuestos, este es el modelo más apropiado para este conjunto de datos. Además, muestra que no hay otras combinaciones de variables explicativas que suministren valores de la estadística AIC, lo cual confirma que no se encuentran otras alternativas de modelos que incluyan la variables Ttumor e IGleason en un modelo.

Se considera ahora el efecto del tratamiento DES, el cual toma el valor 1 para los pacientes que recibieron el placebo, y 2 para quien es tratado con DES. Cuando la covariable Trat se adiciona al modelo que contiene Ttumor e IGleason, el valor de $-2\ln\widehat{L}$ se reduce a 22.572. Esta reducción (en 0.951) no es significativa e indica que no hay efecto del tratamiento. No obstante, se verifica si hay efecto del tratamiento sobre estas variables. Se

definen la variables $Trat_Ttumor = Trat \times Ttumor$ y $TI = Trat \times IGleason$ y se adicionan estas interacciones al modelo que contiene Ttumor, UGleason y Trat. Cuando $Trat_Ttumor$ y TI se adicionan al modelo, $-2\ln\widehat{L}$ se reduce a 20.829 y 20.792, respectivamente. Cuando se incluyen ambos términos en el modelo, $-2\ln\widehat{L}$ se reduce a 19.705. Esta reducción obtenida al incluir estas dos interacciones en el modelo no es significativa, luego, no hay evidencia de que el efecto del tratamiento dependa del tamaño del tumor y del índice de Gleason. Así, la conclusión es que el tratamiento con DES no parece tener efecto sobre el riesgo de muerte.

6.4.2.3. Método de residuales escalados de Schoenfeld y una función del tiempo

Un método para visualizar el supuesto de riesgos proporcionales propuesto por Therneau consiste en elaborar un gráfico de los residuos escalados de Schoenfeld (6.4.2) frente a una función del tiempo $g(t)$. Una inclinación de cero de la gráfica (horizontal) advierte sobre la proporcionalidad de los riesgos. Una ayuda gráfica para la detección de posibles desvíos en el supuesto de proporcionalidad consiste en una banda de confianza sobre la nube de puntos, con la cual se visualiza un poco más fácil la horizontalidad o no del gráfico. Además, se puede calcular el coeficiente de correlación, ρ, de los residuos escalados de Schoenfeld y la función auxiliar $g(t)$ para cada covariable. Los valores de ρ cercanos a cero evidencian que el supuesto de proporcionalidad se mantiene en los datos modelados. Una verificación de la hipótesis $H_0 : \rho = 0$ se desarrolla para cada covariable $j = 1, \ldots, p$ mediante la estadística

$$T_j = \frac{e\left(\sum_k (g_k - \hat{g}) s^*_{jk}\right)^2}{\mathcal{I}O_j^{-1} \sum_k (g_k - \hat{g})^2}, \tag{6.4.6}$$

donde e es el número de eventos y $\sum_k (g_k - \hat{g})$ es el j-ésimo elemento de la diagonal del inverso de la matriz de información observada. La variable aleatoria T_j tiene, aproximadamente, una distribución ji-cuadrado con un grado de libertad. Los valores de $T_j > \chi^2_{1,1-\alpha}$ evidencian que el supuesto de riesgos proporcionales no se cumple para la j-ésima covariable.

En el paquete R se encuentran disponibles algunas opciones para la forma funcional de $g(t)$, tales como t, $\ln(t)$, rank (rango) y km (la función de sobrevida de Kaplan-Meier continua a la izquierda).

Una verificación global de la proporcionalidad de los riesgos, que involucra todas las covariables incluidas en el modelo y considera $g_j(t) = g(t)$, se hace mediante la estadística

$$T = \frac{(g - \bar{g})' S^* (\mathcal{I}O) S^{*'} (g - \bar{g})}{e \sum_k (g_k - \hat{g})^2}, \tag{6.4.7}$$

donde $\mathcal{I}O$ es la matriz de información observada, e es el número de eventos y $S^* = e\mathcal{R}\mathcal{I}O^{-1}$, con $\mathcal{R}$ como una matriz $e \times p$ de los residuos de Schoenfeld. Esta estadística tiene una distribución ji-cuadrado con p grados de libertad. Los valores p de la estadística (6.4.7) menores que un valor α, especificado previamente, evidencian el incumplimiento del supuesto de proporcionalidad de los riesgos al evento en el modelo ajustado para los datos.

6.4.2.4. Métodos con covariable dependiente del tiempo

Una alternativa para evaluar el supuesto de riesgos proporcionales en el modelo de Cox consiste en incluir en el modelo una covariable dependiente del tiempo. El modelo de Cox con covariables dependientes del tiempo extienden este modelo como se muestra en la ecuación (6.2.4).

Para simplificar, suponga que se tiene como covariable un factor que indica la pertenencia a uno de los dos grupos (por ejemplo, medicamento y placebo). Se trata de verificar si la razón de tasas del evento es la misma en cualquier tiempo t. Nuevamente, el modelo de Cox para este caso es dado por

$$h(t|x_1) = h_0(t) \exp\{x_1 \beta_1\},$$

con x_1 siendo la covariable que indica la pertenencia o no a un grupo de interés. Como se ha mostrado anteriormente, si el modelo es adecuado, se tiene que la razón de las tasas de riesgos de un grupo respecto al otro es $\exp\{\beta_1\}$.

Otra covariable $x_2 = t$ puede ser adicionada al modelo, esto se hace de la forma

$$h(t|x_1) = h_0(t) \exp\{x_1 \beta_1 + \beta_2 t\}.$$

La razón de tasas de falla, entre el riesgo evaluado en $x_1 = 1$ y el riesgo evaluado en $x_1 = 0$ es, entonces,

$$\exp\{\beta_1 + \beta_2 t\}.$$

Esta razón no es constante en el tiempo y, por lo tanto, el modelo no es de riesgos proporcionales. Si $\beta_2 < 0$, entonces la razón de las tasas de riesgos decrece con el tiempo. Esto significa que el riesgo al evento disminuye con el tiempo para el grupo de interés. Por otra parte, si $\beta_2 > 0$, el riesgo al evento aumentaría con el tiempo para el grupo de interés. En el caso de $\beta_2 = 0$, el riesgo es constante e igual a $\exp\{\beta_1\}$, con lo cual se evidencia que esta

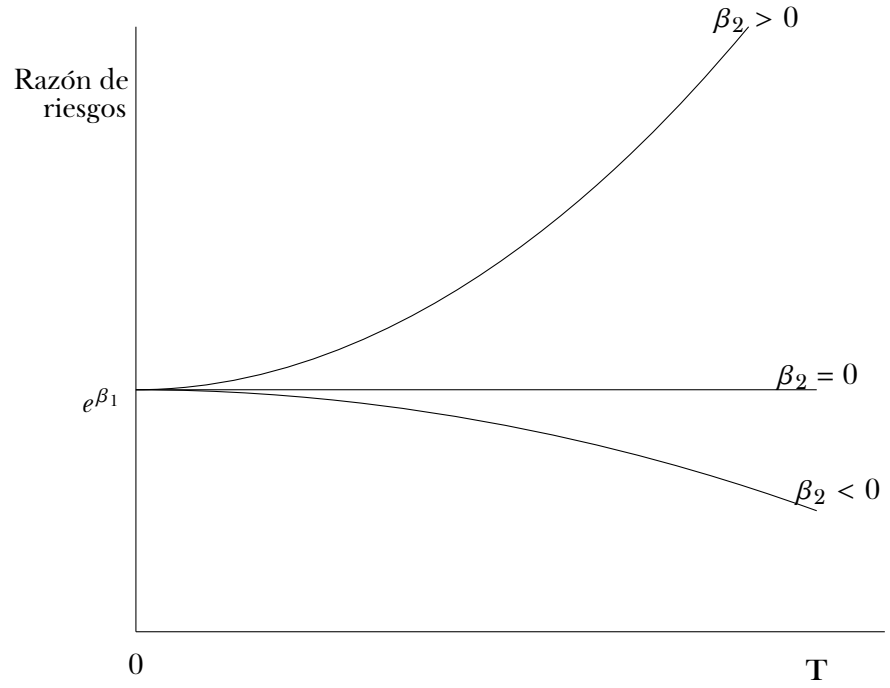

Figura 6.3. Razón de tasas de falla $\exp\{\beta_1 + \beta_2 t\}$.

hipótesis corresponde al supuesto de riesgos proporcionales. La figura 6.3 muestra los diferentes escenarios del efecto sobre el riesgo al evento de la variable dependiente del tiempo. El ajuste al modelo de Cox, en términos de covariables que dependen del tiempo, no es el mismo que en el caso de las covariables invariantes en el tiempo. Una razón simple al observar las ecuaciones de estimación (6.3.6) es que, como las covariables toman valores diferentes en cada punto del tiempo, esto haría más complejo los cálculos del denominador. De todas maneras, la estrategia gráfica con los residuos es una herramienta apropiada para diagnosticar el modelo.

6.4.2.5. Puntos atípicos, influyentes y la forma funcional de las covariables

Los residuos de Martingala definidos en (6.4.3), es decir,

$$r_{Mi} = \delta_i - r_{CSi} = \delta_i - \widehat{H}_0(t) \exp\left\{\sum_{k=1}^{p} x_{ik}\hat{\beta}_k\right\}, \ i = 1, \ldots, n,$$

se emplean para verificar tanto la presencia de datos atípicos como la forma funcional (en potencia, logaritmo, entre otras) de cómo debe incluirse

las covariables en el modelo. Por ejemplo, con una variable continua x_j, la gráfica de r_{Mi} frente a x_{ij} se emplea para diagnosticar la forma funcional de esta covariable. La interpretación de estas gráficas no es tan inmediata y requiere un poco de "ensayo error", algo de paciencia y, sobre todo, experticia en el marco conceptual de donde proceden los datos.

La *deviance* es otro residuo que se puede emplear para advertir sobre datos atípicos. Para el modelo de Cox, de acuerdo con la definición de este residuo dada en (6.4.4), se expresa como

$$r_{Di} = sign(r_{Mi})\sqrt{-2[r_{Mi} + \delta_i \ln(\delta_i - r_{Mi})]}.$$

Estos residuos son más simétricos que los de Martingala. Un gráfico de r_{Di} frente al predictor lineal $\sum_{j=1}^{p} x_{ij}\hat{\beta}_j$, para cada unidad $i = 1, \ldots, n$ y cada variable $j = 1, \ldots, p$, es empleado para detectar datos atípicos en cada unidad y cada variable.

Otro uso de los residuales es la evaluación de cada observación sobre su influencia en el modelo. Una estrategia consiste en comparar una estadística con y sin la observación para determinar la "sensibilidad" del modelo a cada observación, es decir, el impacto o la influencia de cada observación sobre el modelo. La medida usual es el residuo de JackKnife dado por

$$J_i = \hat{\beta} - \hat{\beta}_{(i)}, \ i = 1, \ldots, n,$$

donde $\hat{\beta}_{(i)}$ es la estimación del parámetro β sin la i-ésima observación. La influencia de cada observación, $J - i$, es proporcional a $(x_i - \bar{x}) * residuo$. De esta manera, una observación puede tener una influencia grande, ya sea por estar muy distanciada del promedio o por tener un residuo grande.

Para el modelo de Cox, la influencia de cada observación se hace considerando tanto la distancia de la observación a la media o como la discrepancia de la observación con el modelo (residuo). La estadística que incluye los dos tipos de influencia es

$$\Delta\beta_i = (\hat{\beta} - \hat{\beta}_{(i)})'[\mathcal{I}O(\hat{\beta})](\hat{\beta} - \hat{\beta}_i), \ i = 1, \ldots, n,$$

con $\Delta\beta$ siendo el cambio del vector de coeficientes estimados obtenido por la exclusión de una observación en cada etapa $i = 1, \ldots, n$. Un gráfico de $\Delta\beta_i$ frente a i puede advertir sobre observaciones influyentes. En la literatura del análisis de regresión, esta estadística se conoce como la distancia de Cook.

Esta estadística se puede obtener para cada covariable incluida en el modelo de Cox. Así, la estadística para cada unidad en cada variable es

$$\Delta\beta_{ij} = (\hat{\beta}_j - \hat{\beta}_{j,(i)})'[\mathcal{I}O(\hat{\beta})]_{jj}(\hat{\beta}_j - \hat{\beta}_{j,(i)}), \ i = 1, \ldots, n \text{ y } j = 1, \ldots, p.$$

Estas distancias conforman una matriz D de tamaño $n \times p$, denominada la matriz de los residuos *dfbetas*. Como en el análisis de modelos de regresión clásicos, los gráficos de los residuos *dfbetas* ligados con cada covariable, frente a los valores de las covariables, se emplean para explorar las observaciones influyentes.

Ejemplo 6.4.2. Se toman los datos de un estudio sobre 90 pacientes de sexo masculino diagnosticados con cáncer de laringe, los cuales se reportan en Klein y Moeschberger [39] y en Colosimo y Giolo [11]. Para cada paciente se registró su edad (en años) y el estadio (*est*) de la enfermedad que corresponde a:

- I: tumor primario,
- II: envoltura de nódulos,
- III: metástasis,
- IV: combinación de los tres estadios anteriores.

Además, se registró el tiempo al evento muerte o censura (en meses). Por el grado de la severidad de la enfermedad, los estadios son de tipo ordinal.

Para el análisis de estos datos se emplea el modelo de riesgos proporcionales de Cox. El procesamiento (cálculo y gráficas) se elaboran con el paquete R; a continuación se escribe el respectivo código o sintaxis de R para cada uno de los cuatro modelos ajustados (cox_lar1 a cox_lar4).

La expresión cox_lar1 corresponde al modelo de Cox sin covariables (modelo nulo). El código R para hacer los cálculos de estimación es el siguiente.

```
laringe<-read.table("Laringe.txt",header=T) # datos tabla 6.7
attach(laringe)
require(survival)
require(MASS)
cox_lar1<-coxph(Surv(tpo,cens)~factor(est),data=laringe,x=T,
+method="breslow")
summary(cox_lar1)
cox_lar1$loglik
```

El nombre cox_lar2 corresponde al modelo de Cox con la covariable estadio (*est*), la cual se considera como un factor con cuatro niveles; para este factor se crean tres variables ficticias guiadas por la tabla 6.1. El código R para hacer los cálculos de estimación es el siguiente.

```
cox_lar2<-coxph(Surv(tpo,cens)~factor(est),data=laringe,x=T,
+method="breslow")
summary(cox_lar2)
cox_lar2$loglik
```

Al modelo de Cox anterior se le adiciona la variable *edad*, que es denominada cox_lar3. El código R para estimar el modelo y calcular el logaritmo de la razón de verosimilitud es como sigue.

```
cox_lar3<-coxph(Surv(tpo,cens)~factor(est)+edad,data=laringe,
+x=T,method="breslow")
summary(cox_lar3)
cox_lar3$loglik
```

Finalmente, el modelo de Cox que incluye *estadio*, *edad* y la interacción *estadio* y *edad* es nominado como cox_lar4. El código R para procesar este modelo es el siguiente.

```
cox_lar4<-coxph(Surv(tpo,cens)~factor(est)+edad+
+factor(est)*edad,data=laringe,x=T,method="breslow")
summary(cox_lar4)
cox_lar4$loglik
```

En la tabla 6.8 se editan los resultados de este procesamiento. Desde esta tabla se tiene $-2\ln\widehat{L} = -2 * (185.0775 - 188.1794) = 6.2038$. El valor p para una distribución ji-cuadrado con 3 grados de libertad es $p = 0.1021051$, con lo cual se puede concluir que la interacción no es significativa. No obstante, las verificaciones individuales sobre los parámetros de esta interacción evidencian que al menos uno de los β's asociados con la interacción difiere significativamente de cero. Se trata de $\hat{\beta}_5$ con un valor $p = 0.0215$.

Los resultados sobre la estimación, error estándar y verificación de la hipótesis de no efecto sobre los parámetros de interacción *edad* $*$ *estadio* están contenidos en la tabla 6.9. Un complemento al análisis del ajuste de los modelos de Cox propuestos para los datos de cáncer de laringe son los residuos escalados de Schoenfeld (6.4.2). Se hace el análisis de residuales para los modelos de Cox con y sin interacción y a partir de este análisis se debe decidir cuál de los dos modelos se ajusta más adecuadamente a los datos.

Los residuos escalados de Schoenfeld para el modelo de Cox con interacción se obtienen mediante el código R como se muestra a continuación.

```
residuals(cox_lar4,type="scaledsch")
cox.zph(cox_lar4,transform="identity") ## g(t)=t
par(mfrow=c(2,4))
plot(cox.zph(cox_lar4))
```

De los resultados de la tabla 6.10, se advierte que las correlaciones, tipo Pearson (ρ), estimadas entre los residuales de Schoenfeld y la función auxiliar $g(t)$, son cercanas a cero para cada variable. Además, no se rechaza la hipótesis $H_0: \rho = 0$, pues los valores p asociados con la estadística (6.4.6) son mayores a 0.33, con lo cual se respalda, desde los datos, el supuesto de proporcionalidad de los datos en el modelo ajustado de Cox.

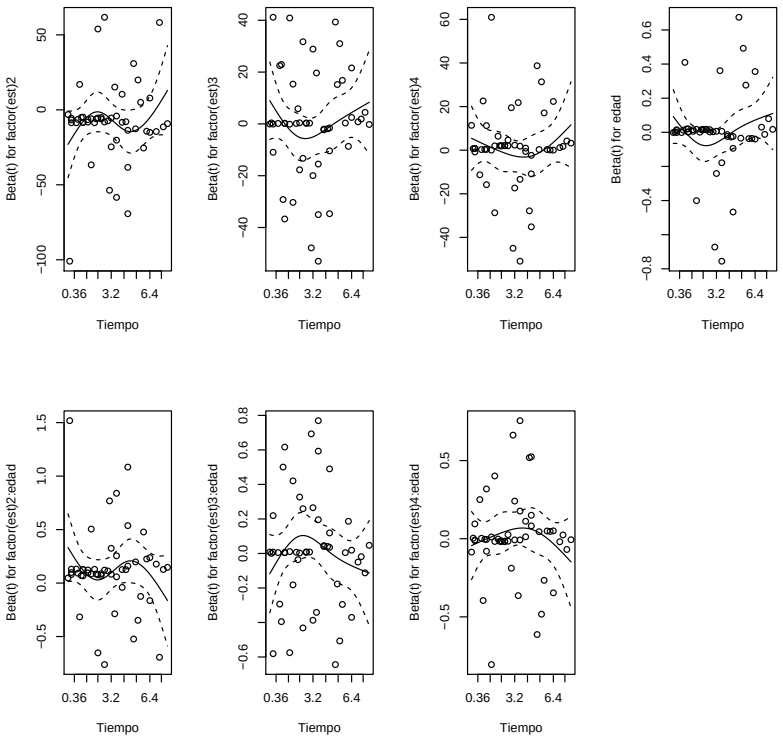

Figura 6.4. Residuos de Schoenfeld frente al tiempo por *estadio* y *edad*.

A partir de los gráficos de los residuos de Schoenfeld versus $g(t) = t$ que se muestran en la figura 6.4, obtenidos con el código R anterior, se ve que no hay tendencias alejadas de la horizontal. En resumen, las tres herramientas estadísticas, a saber, $\hat{\rho}$, verificación de $H_0: \rho = 0$ y las gráficas de los

residuales, coinciden en advertir que estos datos no muestran evidencias suficientes para rechazar el supuesto de proporcionalidad.

Así, el modelo de Cox cuyo predictor lineal contiene las variables *estadio*, *edad* y la interacción *estadio* $*$ *edad* se ajusta adecuadamente a los datos contenidos en la tabla 6.8.

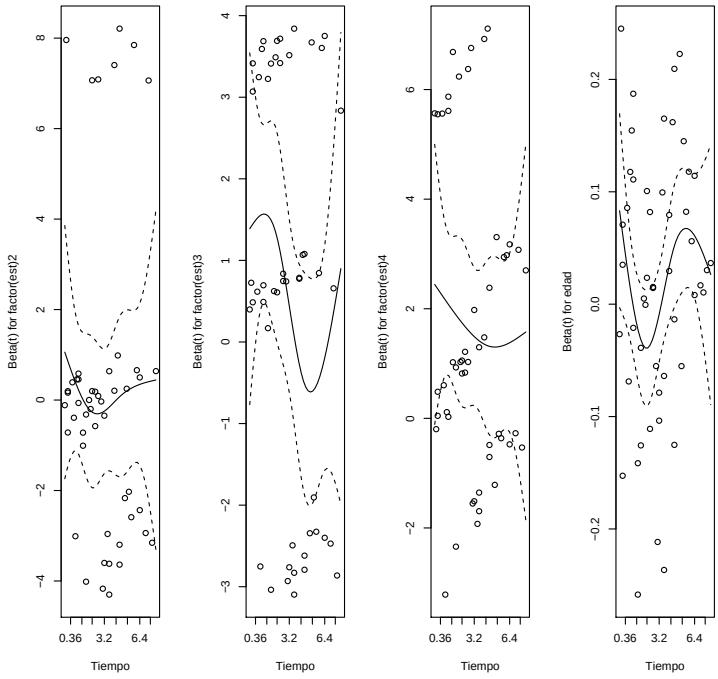

Figura 6.5. Residuos de Schoenfeld frente al tiempo por *estadio* y *edad*.

Un análisis similar al anterior, aplicado a los residuos del modelo sin el término de interacción, se resume en las estadísticas del coeficiente de correlación de Pearson (ρ) entre los residuos de Schoenfeld y la función auxiliar $g(t)$. La verificación $H_0: \ \rho = 0$ y el respectivo valor p se escriben en la tabla 6.11.

A partir de la tabla 6.11, se observa que el estadio III puede resultar significativo, con lo cual se advierte sobre un posible incumplimiento del supuesto de riesgos proporcionales para este nivel de la covariable *estadio*.

Las gráficas de residuales contenidas en la figura 6.5, excepto la asociada con el estadio III, muestran una tendencia casi horizontal de las nubes de puntos, por lo que se puede afirmar el cumplimiento del supuesto de proporcionalidad.

Del análisis y diagnóstico de los dos modelos de Cox ajustados para los datos de cáncer de laringe, se recomienda el modelo de Cox con presencia del término de interacción. La tabla 6.12 muestra las estimaciones de los respectivos parámetros, junto con el error estándar, el valor p de la estadística asociada con la hipótesis nula $H_0 : \beta_j = 0$ y la razón de riesgo (RR).

Considere las estimaciones obtenidas para los parámetros del modelo seleccionado. Sea $\mathbf{x} = (estadio, edad)$, la estimación de las funciones de sobrevida y de riesgo acumulado, de acuerdo con los datos de la tabla 6.12, son dadas, respectivamente, por

$$\widehat{S}(t|\mathbf{x}) = \begin{cases} \widehat{S}_0(t)^{\exp\{\hat{\beta}_4 edad\}}, & \text{si estadio I} \\ \widehat{S}_0(t)^{\exp\{\hat{\beta}_1+(\hat{\beta}_4+\hat{\beta}_5)edad\}}, & \text{si estadio II} \\ \widehat{S}_0(t)^{\exp\{\hat{\beta}_2+(\hat{\beta}_4+\hat{\beta}_6)edad\}}, & \text{si estadio III} \\ \widehat{S}_0(t)^{\exp\{\hat{\beta}_3+(\hat{\beta}_4+\hat{\beta}_7)edad\}}, & \text{si estadio IV} \end{cases}$$

y

$$\widehat{h}(t|\mathbf{x}) = \begin{cases} \widehat{h}_0(t)\exp\{\hat{\beta}_4 edad\}, & \text{si estadio I} \\ \widehat{h}_0(t)\exp\{\hat{\beta}_1 + (\hat{\beta}_4 + \hat{\beta}_5)edad\}, & \text{si estadio II} \\ \widehat{h}_0(t)\exp\{\hat{\beta}_2 + (\hat{\beta}_4 + \hat{\beta}_6)edad\}, & \text{si estadio III} \\ \widehat{h}_0(t)^{\exp\{\hat{\beta}_3+(\hat{\beta}_4+\hat{\beta}_7)edad\}}, & \text{si estadio IV.} \end{cases}$$

Para efectos de lectura gráfica y, por tanto, de interpretación, resultan más cómodas las gráficas de la función de riesgo acumulado. De acuerdo con los valores de $\mathbf{x}$, estas vienen dadas por

$$\widehat{H}(t|\mathbf{x}) = \widehat{H}(t)_0 \exp\{\mathbf{x}'\hat{\beta}\}.$$

Con el modelo ajustado a los datos se puede estimar tanto la función de sobrevida, $\widehat{S}(t|\mathbf{x})$, como la función de riesgo acumulado, $\widehat{H}(t|\mathbf{x})$. El código R para calcular la estimación de la función de sobrevida y de riesgo es:

```
H<-basehaz(cox_lar4,centered=F)
tpos<-H$time
H0<-H$hazard
S0<-exp(-H0)
round(cbind(tpos,S0,H0),digits=5)
```

La figura 6.6 contiene las gráficas de las funciones de sobrevida estimadas para las edades de 50 y 65 años de pacientes con cáncer de laringe, en cada uno de los cuatro estadios de la enfermedad. Estas se generan mediante el siguiente código del paquete R.

```
par(mfrow=c(1,2))
# GRAFICAS DE SOBREVIDA 50 A\~nOS
plot(tpos,S0^(exp(-0.002559*50)),ylim=c(0,1),type="l",
+xlab="Tiempo",ylab="S(t|x)")
lines(tpos,S0^(exp(-7.946142+(-0.002559+0.120254)*50)), lty=2)
lines(tpos,S0^(exp(-0.122500+(-0.002559+0.011351)*50)), lty=3)
lines(tpos,S0^(exp(0.846986+(-0.002559+0.013673)*50)), lty=4)

# GRAFICAS DE SOBREVIDA 65 A\~nOS
plot(tpos,S0^(exp(-0.002559*65)),ylim=c(0,1),type="l",
+xlab="Tiempo",ylab="S(t|x)")
lines(tpos,S0^(exp(-7.946142+(-0.002559+0.120254)*65)), lty=2)
lines(tpos,S0^(exp(-0.122500+(-0.002559+0.011351)*65)), lty=3)
lines(tpos,S0^(exp(0.846986+(-0.002559+0.013673)*65)), lty=4)
```

En ambos grupos de edad el estadio II tiene mayor probabilidad de sobrevida. Sin embargo, en el grupo de 65 años, la probabilidad de sobrevida de los pacientes en este estadio decrece notoriamente y se asemeja a la curva de sobrevida de los pacientes en estadio I. En ambos grupos de edad los pacientes cuyo estadio de la enfermedad es IV tienen, relativamente, menor probabilidad de sobrevida. En conjunto, el grupo de pacientes con 65 años de edad tiene una sobrevida menor (en términos de probabilidad) que los pacientes con 50 años de edad. Esta característica avala la inclusión del término de interacción.

La figura 6.7 contiene las gráficas ligadas a las funciones de riesgo acumulado estimadas para pacientes con cáncer de laringe en los grupos de edad de 50 y 65 años, considerando cada uno de los cuatro estadios de la enfermedad. El siguiente es el código del paquete R con el cual se generaron estas gráficas.

```
par(mfrow=c(1,2))
plot(tpos,H0*(exp(-0.002559*50)),xlim=c(0,8),ylim=c(0,4),
+type="l",lty=1,xlab="Tiempo",main="",ylab="Riesgo Acumulado:
+ H(t|x)")
legend("topleft",lty=c(1,2,3,4),c("Estadio I",
+"Estadio II","Estadio III","Estadio IV"),bty="n",cex=0.6)
title(main=list("Edad: 50 a\~nos"), cex=0.5)
lines(tpos,H0*(exp(-7.946142+(-0.002559+0.120254)*50)), lty=2)
lines(tpos,H0*(exp(-0.122500+(-0.002559+0.011351)*50)), lty=3)
lines(tpos,H0*(exp(0.846986+(-0.002559+0.013673)*50)), lty=4)
```

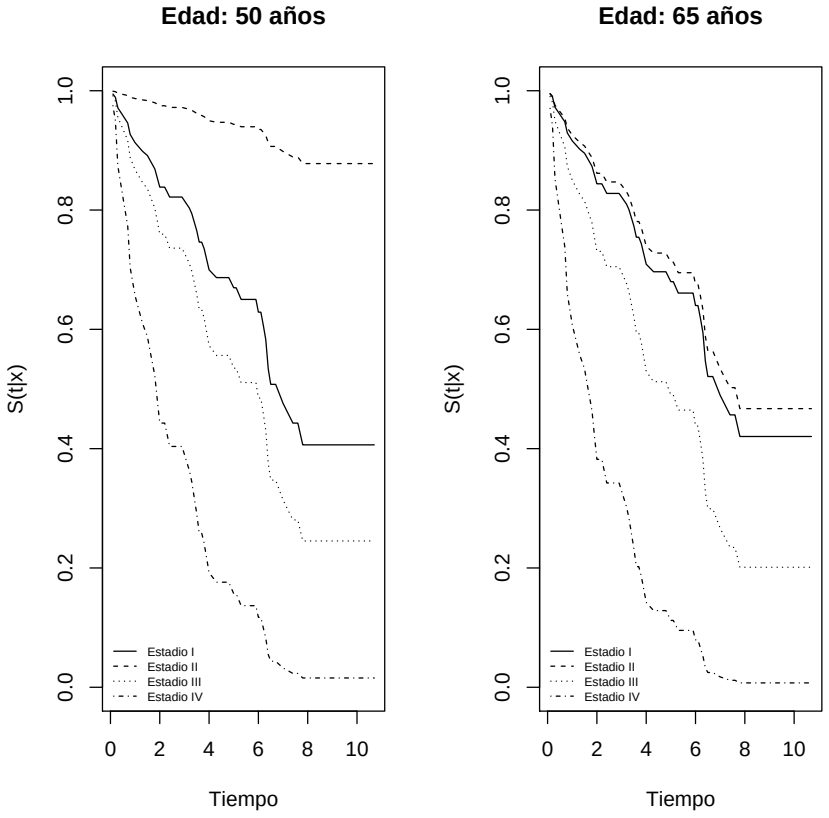

Figura 6.6. Funciones de sobrevida estimadas por el modelo de Cox para los datos de cáncer de laringe.

```
#### GRAFICAS 65
plot(tpos,H0*(exp(-0.002559*65)),xlim=c(0,8),ylim=c(0,4),
+type="l",lty=1, xlab="Tiempo",main="",ylab="Riesgo Acumulado:
+H(t|x)")
title(main=list("Edad: 65 a\~nos"), cex=0.5)
legend("topleft",lty=c(1,2,3,4),c("Estadio I","Estadio II",
+"Estadio III","Estadio IV"),bty="n",cex=0.6)
lines(tpos,H0*(exp(-7.946142+(-0.002559+0.120254)*65)), lty=2)
lines(tpos,H0*(exp(-0.122500+(-0.002559+0.011351)*65)), lty=3)
lines(tpos,H0*(exp(0.846986+(-0.002559+0.013673)*65)), lty=4)
```

A partir de estos resultados, se puede hacer una razón de riesgo entre diferentes perfiles de pacientes. Así, por ejemplo, para los pacientes i e i'

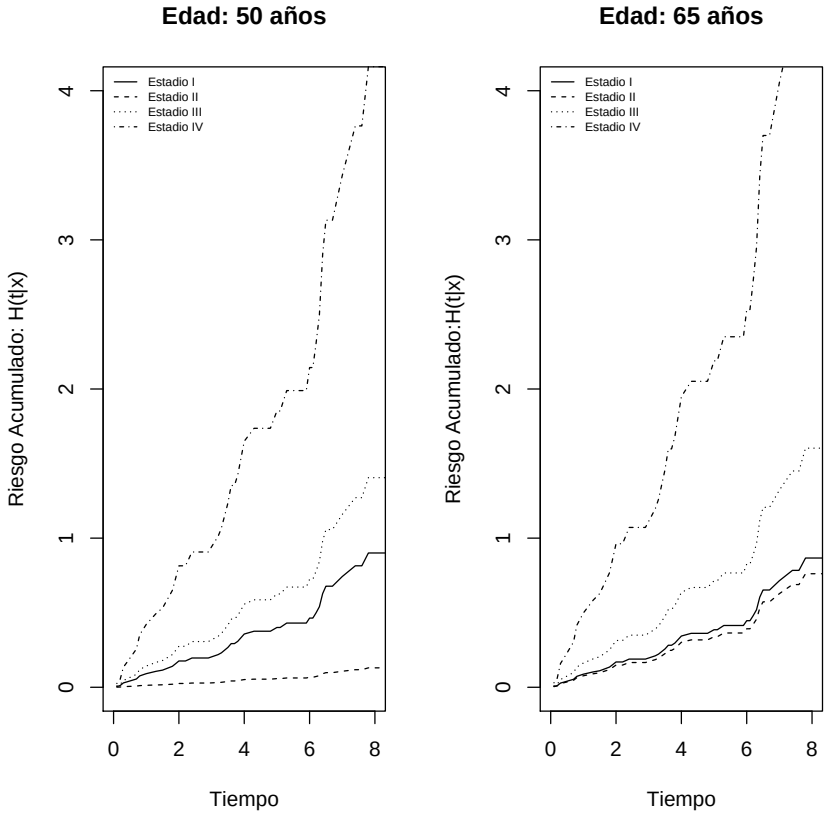

Figura 6.7. Riesgos acumulados estimados por el modelo de Cox para los datos de cáncer de laringe.

que pertenecen al estadio II de la enfermedad y tienen edades de 65 y 50 años, respectivamente, la razón de sus riesgos al evento muerte es

$$\begin{aligned}\frac{\widehat{h}(t|\mathbf{x}_i)}{\widehat{h}(t|\mathbf{x}_{i'})} &= \frac{\exp\{\hat{\beta}_1 + (\hat{\beta}_4 + \hat{\beta}_5) * 65\}}{\exp\{\hat{\beta}_1 + (\hat{\beta}_4 + \hat{\beta}_5) * 50\}} \\ &= \exp\left\{(\hat{\beta}_4 + \hat{\beta}_5) * (65 - 50)\right\} \\ &= 5.84.\end{aligned}$$

Este resultado indica que los pacientes con estadio de enfermedad II y con 65 años de edad tienen un riesgo de muerte 6 veces mayor con respecto a los pacientes que tienen 50 años de edad y están en el mismo estadio de la enfermedad.

Ahora se hace una comparación entre pacientes de la misma edad y del mismo estadio de la enfermedad. Por ejemplo, los pacientes de 50 años de edad clasificados en estadio de enfermedad IV y III, respectivamente. Sean j y j' tales pacientes, la razón de riesgos es

$$\frac{\widehat{h}(t|\mathbf{x}_j)}{\widehat{h}(t|\mathbf{x}_{j'})} = \frac{\exp\{\hat{\beta}_3 + (\hat{\beta}_4 + \hat{\beta}_7) * 50\}}{\exp\{\hat{\beta}_2 + (\hat{\beta}_4 + \hat{\beta}_6) * 50\}} = 2.96.$$

Así, el riesgo al evento muerte de los pacientes con 50 años en el estadio de cáncer de laringe tipo IV es aproximadamente 3 veces el riesgo de muerte de los pacientes con la misma edad en el estadio III.

Tabla 6.7. Datos de pacientes con cáncer de laringe

Id	tpo	cens	edad	est	Id	tpo	cens	edad	est
1	0.6	1	77	1	46	4.3	0	64	2
2	1.3	1	53	1	47	5.0	0	66	2
3	2.4	1	45	1	48	7.5	0	50	2
4	3.2	1	58	1	49	7.6	0	53	2
5	3.3	1	76	1	50	9.3	0	61	2
6	3.5	1	43	1	51	0.3	1	49	3
7	3.5	1	60	1	52	0.3	1	71	3
8	4.0	1	52	1	53	0.5	1	57	3
9	4.0	1	63	1	54	0.7	1	79	3
10	4.3	1	86	1	55	0.8	1	82	3
11	5.3	1	81	1	56	1.0	1	49	3
12	6.0	1	75	1	57	1.3	1	60	3
13	6.4	1	77	1	58	1.6	1	64	3
14	6.5	1	67	1	59	1.8	1	74	3
15	7.4	1	68	1	60	1.9	1	53	3
16	2.5	0	57	1	61	1.9	1	72	3
17	3.2	0	51	1	62	3.2	1	54	3
18	3.3	0	63	1	63	3.5	1	81	3
19	4.5	0	48	1	64	5.0	1	59	3
20	4.5	0	68	1	65	6.3	1	70	3
21	5.5	0	70	1	66	6.4	1	65	3
22	5.9	0	47	1	67	7.8	1	68	3
23	5.9	0	58	1	68	3.7	0	52	3
24	6.1	0	77	1	69	4.5	0	66	3
25	6.2	0	64	1	70	4.8	0	54	3
26	6.5	0	79	1	71	4.8	0	63	3
27	6.7	0	61	1	72	5.0	0	49	3
28	7.0	0	66	1	73	5.1	0	69	3
29	7.4	0	73	1	74	6.5	0	65	3
30	8.1	0	56	1	75	8.0	0	78	3
31	8.1	0	73	1	76	9.3	0	69	3
32	9.6	0	58	1	77	10.1	0	51	3
33	10.7	0	68	1	78	0.1	1	65	4
34	0.2	1	86	2	79	0.3	1	71	4
35	1.8	1	64	2	80	0.4	1	76	4
36	2.0	1	63	2	81	0.8	1	65	4
37	3.6	1	70	2	82	0.8	1	78	4
38	4.0	1	81	2	83	1.0	1	41	4
39	6.2	1	74	2	84	1.5	1	68	4
40	7.0	1	62	2	85	2.0	1	69	4
41	2.2	0	71	2	86	2.3	1	62	4
42	2.6	0	67	2	87	3.6	1	71	4
43	3.3	0	51	2	88	3.8	1	84	4
44	3.6	0	72	2	89	2.9	0	74	4
45	4.3	0	47	2	90	4.3	0	48	4

Tabla 6.8. Modelos de Cox para datos de pacientes con cáncer de laringe

Modelo	Covariables	Estimación	log verosim. parcial
1	Ninguna	-	l_1=-197.2129
2	est: Estadio II	$\hat{\beta}_1 = 0.06576$	
	Estadio III	$\hat{\beta}_2 = 0.61206$	
	Estadio IV	$\hat{\beta}_3 = 1.72284$	l_2=-189.0812
3	est: Estadio II	$\hat{\beta}_1 = 0.13856$	
	Estadio III	$\hat{\beta}_2 = 0.63835$	
	Estadio IV	$\hat{\beta}_3 = 1.69306$	
	edad	$\hat{\beta}_4 = 0.01890$	l_3=-188.1794
4	est: Estadio II	$\hat{\beta}_1 = -7.946142$	
	Estadio III	$\hat{\beta}_2 = -0.122500$	
	Estadio IV	$\hat{\beta}_3 = 0.846986$	
	edad	$\hat{\beta}_4 = -0.002559$	
	est*edad: (II*edad)	$\hat{\beta}_5 = 0.120254$	
	(III*edad)	$\hat{\beta}_6 = 0.011351$	
	(IV*edad)	$\hat{\beta}_7 = 0.013673$	l_4=-185.0775

Tabla 6.9. Verificación de modelos con parámetros de interacción

Parámtetro	Estimación	Error estándar	Wald	Valor p
β_5: edad & estadio II	0.120254	0.052307	2.2990	0.0215
β_6: edad & estadio III	0.011351	0.037449	0.3031	0.7618
β_7: edad & estadio IV	0.013673	0.035967	0.3802	0.7038

Tabla 6.10. Verificación de la proporcionalidad en el modelo de Cox con interacción *estadio* y *edad*

Covariable	Correlación (ρ)	ji-cuadrado (χ^2)	Valor p
Estadio II	0.0958	0.5033	0.478
Estadio III	0.0462	0.1577	0.691
Estadio IV	0.0269	0.0421	0.837
Edad	0.1082	0.9376	0.333
Estadio II * Edad	-0.0943	0.4929	0.483
Estadio III * Edad	-0.0768	0.4364	0.509
Estadio IV * Edad	-0.0443	0.1160	0.733
Tendencia global	-	5.7988	0.563

Tabla 6.11. Verificación de la proporcionalidad en el modelo de Cox sin la interacción *estadio* y *edad*

Covariable	Correlación (ρ)	ji-cuadrado (χ^2)	Valor p
Estadio II	-0.0107	0.00605	0.9380
Estadio III	-0.2440	2.83791	0.0921
Estadio IV	-0.1188	0.62202	0.4303
Edad	0.1328	1.16886	0.2796
Tendencia global	-	4.56330	0.3351

Tabla 6.12. Modelo de regresión de Cox ajustado para los datos de cáncer de laringe

Covariable	Estimación	Razón Ries.	Error Est.	Z	Valor p
X_j	$\hat{\beta}_j$	$\exp(\hat{\beta}_j)$	$ee(\hat{\beta}_j)$		$P(\|Z\| > z)$
Estadio II	-7.946142	0.000354	3.678209	-2.160	0.0307
Estadio III	-0.122500	0.884706	2.468331	-0.050	0.9604
Estadio IV	0.846986	2.332605	2.425717	0.349	0.7270
Edad	-0.002559	0.997444	0.026051	-0.098	0.9218
Estadio II * Edad	0.120254	1.127783	0.052307	2.299	0.0215
Estadio III * Edad	0.011351	1.011416	0.037449	0.303	0.7618
Estadio IV * Edad	0.013673	1.013767	0.035967	0.380	0.7038

Tabla 6.13. Estimación de $\hat{S}_0(t)$ y $\hat{H}_0(t)$ para datos de cáncer de laringe

ID	Tiempo	$\hat{S}_0(t)$	$\hat{H}_0(t)$	ID	Tiempo	$\hat{S}_0(t)$	$\hat{H}_0(t)$
1	0.1	0.99377	0.00625	28	4.0	0.66650	0.40572
2	0.2	0.98739	0.01269	29	4.3	0.65242	0.42707
3	0.3	0.96737	0.03318	30	4.5	0.65242	0.42707
4	0.4	0.96039	0.04041	31	4.8	0.65242	0.42707
5	0.5	0.95319	0.04794	32	5.0	0.63406	0.45561
6	0.6	0.94596	0.05555	33	5.1	0.63406	0.45561
7	0.7	0.93875	0.06321	34	5.3	0.61308	0.48925
8	0.8	0.91713	0.08650	35	5.5	0.61308	0.48925
9	1.0	0.90154	0.1036	36	5.	0.61308	0.48925
10	1.3	0.88552	0.12158	37	6.0	0.59024	0.52723
11	1.5	0.87745	0.13073	38	6.1	0.59024	0.52723
12	1.6	0.86907	0.14033	39	6.2	0.56680	0.56775
13	1.8	0.85231	0.15980	40	6.3	0.54126	0.61386
14	1.9	0.83549	0.17974	41	6.4	0.48988	0.71360
15	2.0	0.81848	0.20030	42	6.5	0.46291	0.77022
16	2.2	0.81848	0.20030	43	6.7	0.46291	0.77022
17	2.3	0.80945	0.21140	44	7.0	0.43005	0.84384
18	2.4	0.80004	0.22309	45	7.4	0.39625	0.92571
19	2.5	0.80004	0.22309	46	7.5	0.39625	0.92571
20	2.6	0.80004	0.22309	47	7.6	0.39625	0.92571
21	2.9	0.80004	0.22309	48	7.8	0.35937	1.02341
22	3.2	0.77965	0.24891	49	8.0	0.35937	1.02341
23	3.3	0.76923	0.26236	50	8.1	0.35937	1.02341
24	3.5	0.73805	0.30374	51	9.3	0.35937	1.02341
25	3.6	0.71695	0.33274	52	9.6	0.35937	1.02341
26	3.7	0.71695	0.33274	53	10.1	0.35937	1.02341
27	3.8	0.70498	0.34959	54	10.7	0.35937	1.02341

Capítulo

siete

Extensiones del modelo de Cox

$$l(\psi) = \sum_{i=1}^{n} c_i \cdot \exp\left(\frac{x_i'(t)\psi}{w}\right) - \ln\left(\sum_{l \in R_i} \exp\left(\frac{x'(l)(t)\psi}{w}\right)\right)$$

7.1. Introducción

La verosimilitud parcial para un conjunto de datos referentes al tiempo hasta la ocurrencia de un evento permite modelar estos datos junto con la información de las covariables. En este contexto, los valores de las covariables se deben determinar en tiempo, t_0. Por ejemplo, cuando el paciente ingresa al estudio, estos valores permanecen constantes durante el mismo. Dado que en varias situaciones los datos evolucionan o cambian durante el tiempo de estudio, no es adecuado utilizar únicamente el valor inicial de las covariables para modelar información que es dinámica en el tiempo. Para considerar e incluir los valores de las covariables que pueden cambiar su valor en el transcurso del tiempo ("dependientes del tiempo"), se requieren técnicas apropiadas para obtener estimaciones válidas de los parámetros ligados a dichas covariables dependientes del tiempo.

Las extensiones del modelo de riesgos proporcionales de Cox se presentan fundamentalmente por la naturaleza de las covariables o los parámetros, la forma funcional del predictor $\mathbf{x}'\beta$ y la forma funcional de la función de riesgo base $h_0(t)$. En algunos estudios, unas covariables pueden tomar valores que cambian durante el curso del estudio. Por ejemplo, en un paciente, pueden cambiar la masa corporal, la presión sanguínea, el colesterol o su cuadro lipídico en el transcurso del estudio. En general, una covariable dependiente del tiempo t se escribe $X_j(t)$, para $j = 1, \ldots, p$. Estas variables pueden incorporarse al modelo de Cox como se muestra en la expresión (6.2.4). En la sección 7.2 se desarrolla el modelo de Cox para este tipo de covariables.

Cuando el supuesto de proporcionalidad no se cumple para alguna covariable, una alternativa para el modelamiento es estratificar o categorizar la variable y desarrollar el modelo de Cox en cada uno de los estratos o categorías de la variable. El supuesto de proporcionalidad en la función de riesgo se debe mantener en cada estrato. En la sección 7.3 se aborda este modelo.

Otra extensión del modelo de Cox es el caso de los datos de tiempo al evento con censura por intervalo. En la sección 7.4 se presenta el modelo de riesgos proporcionales de Cox con tiempos censurados por intervalo.

Un supuesto estructural, a veces no explícito, es que la unidades proceden de un universo homogéneo. Sin embargo, las unidades varían entre sí. Por ejemplo, ante el mismo tratamiento las unidades pueden tener respuestas que varían de una unidad a otra, y lo mismo puede decirse de los valores que toman algunas covariables. Esta heterogeneidad frecuentemente es referida como variabilidad. El problema del modelamiento es considerar e incluir esta variabilidad en un modelo de la variable respuesta *tiempo hasta*

la ocurrencia de un evento. Los denominados modelos de fragilidad o vulnerabilidad, en inglés, *frailty models*, tratan sobre esta situación. La idea central de estos modelos es que los individuos o unidades tienen diferente vulnerabilidad a la ocurrencia del evento, y que a los más vulnerables el evento les ocurrirá antes que a los menos vulnerables o frágiles. Estos modelos se desarrollan en la sección 7.5.

7.2. Modelos con covariables tiempo dependientes

En la mayoría de los modelos de regresión, las variables explicativas o covariables se registran sobre cada unidad en el tiempo origen del estudio. En la mayoría de estudios sobre tiempo hasta la ocurrencia de un evento, las unidades son seguidas durante el tiempo de duración del estudio. Durante este periodo, los valores de ciertas covariables se pueden registrar en algunos puntos de tiempo; tales valores pueden variar sobre la misma variable de un punto del tiempo a otro. Las covariables cuyos valores cambian en el tiempo se conocen como covariables dependientes del tiempo. En esta sección se muestra cómo este tipo de covariables se pueden incorporar al modelo de Cox, y se hace el respectivo análisis de este tipo de datos.

En el modelo de riesgos proporcionales de Cox, la razón de las funciones de riesgo para cualquier par de unidades se asume independiente del tiempo t. No obstante, en un estudio de la variable *tiempo hasta el evento* es frecuente que se incluyan tanto covariables dependientes del tiempo como covariables independientes del tiempo. En este punto es apropiado considerar dos tipos de covariables que cambian sobre el tiempo para una misma unidad, estas se conocen como covariables *internas* y covariables *externas*.

Las covariables internas, asociadas con una unidad en un estudio, son aquellas que solo pueden ser medidas sobre la unidad mientras que el evento no haya ocurrido. Si el evento es la muerte (o una falla), las covariables internas son aquellas que solo pueden observarse cuando el individuo (persona, animal o planta) esté vivo. Por ejemplo, en el caso de las personas, esto pasa con la presión arterial, el colesterol, la capacidad vital, entre otras. Estas covariables reflejan un perfil del individuo y sus valores son asociables con el tiempo de sobrevida de este.

Las covariables externas son covariables tiempo-dependientes que, para conocer su valor, no requieren necesariamente que a la unidad no le haya ocurrido el evento. Si el evento es la muerte, en los seres vivos una variable externa es aquella en la que no es necesario que el individuo esté vivo para

registrar el valor de la covariable externa. Un ejemplo, en las personas, es la edad, pues una vez se conoce esta en el origen del estudio, la edad futura de la persona se puede establecer en cualquier tiempo. Otro ejemplo es la dosis de algunos fármacos, que varía de una manera predeterminada. Un tercer ejemplo es la covariable que existe totalmente independiente de la unidad, tal como la humedad relativa del ambiente, el nivel de dióxido de sulfuro en la atmósfera, entre otras.

De acuerdo con el modelo de Cox presentado en el capítulo 6, la función de riesgo para la unidad i definida por las covariables $\mathbf{x}_i$ es

$$h_i(t \mid \mathbf{x_i(t)}) = h_0(t)\exp(\mathbf{x_i(t)}'\beta) = h_0(t)\exp\left[\sum_{i=1}^{p}\beta_j x_{ij}(t)\right]. \qquad (7.2.1)$$

La función de riesgo base, $h_0(t)$, se interpreta como el riesgo para una unidad en ausencia de covariables.

En este modelo, no necesariamente todas las covariables son dependientes del tiempo; es decir, $x_j(t) = x_j$ para algún $j = 1, \ldots, p$. Se puede separar el conjunto de p covariables en dos conjuntos de covariables; un conjunto, $\mathbf{x_{i1}}$, con p_1 covariables fijas o no dependientes del tiempo, y el otro conjunto con p_2 covariables, $\mathbf{x_{i2}(t)}$, dependientes del tiempo ($p_1 + p_2 = p$). De forma análoga se particiona el vector de parámetros $\beta = (\beta_1, \beta_2)$, donde β_1, de tamaño $p_1 \times 1$, está asociado con las variables no dependientes del tiempo, mientras que β_2, de tamaño $p_2 \times 1$, se asocia con las covariables dependientes del tiempo. Así, el modelo (7.2.1) se escribe como

$$\begin{aligned} h_i(t \mid \mathbf{x_{i1}}, \mathbf{x_{i2}(t)}) &= h_0(t)\exp\{(\mathbf{x_{i1}}, \mathbf{x_{i2}(t)})(\beta_1, \beta_2)'\} \\ &= h_0(t)\exp\left[\sum_{j=1}^{p_1}\beta_{1j}x_{ij} + \sum_{k=1}^{p_2}\beta_{2k}x_{ik}(t)\right]. \end{aligned} \qquad (7.2.2)$$

En el modelo (7.2.1), el riesgo relativo $h_i(t \mid \mathbf{x_i(t)})/h_0(t)$ es también dependiente del tiempo. Esto significa que el riesgo al evento en el tiempo t ya no es proporcional al riesgo base, y el modelo ya no es de riesgos proporcionales.

Para una interpretación de los parámetros β's en este modelo, considere la razón de funciones de riesgo en el tiempo t para dos unidades i e i', con el mismo valor en las covariables fijas o no dependientes del tiempo. Esto es

$$\frac{h_i(t)}{h_{i'}(t)} = \exp\left[\beta_{21}(x_{i1}(t) - x_{i'1}(t) + \cdots + \beta_{2p_2}(x_{ip_2}(t) - x_{i'p_2}(t))\right].$$

Los coeficientes $\beta_k, k = 1, \ldots, p_2$, se pueden interpretar como el logaritmo de la razón de riesgos asociados con dos unidades cuyo valor en la j-ésima covariable en cualquier tiempo t difieren en 1.0, manteniendo el mismo valor en las demás $p - 1$ covariables.

7.2.0.1. Ajuste del modelo de Cox

Para el modelo de regresión de Cox extendido para incorporar covariables dependientes del tiempo, la función log-verosimilitud parcial, a partir de la ecuación (6.3.7), puede generalizarse a

$$l(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \mathbf{x}_i'(t)\beta - \ln \sum_{l \in \mathcal{R}_i} \exp\left(\mathbf{x}_{(l)}'(t)\beta\right) \right\}, \tag{7.2.3}$$

en la cual $\mathcal{R}_i$ es el conjunto de unidades en riesgo al evento en el momento t_i, es decir, el tiempo al evento de la unidad i en el estudio, $i = 1, \ldots, n$. Además, δ_i es un indicador del evento que es 0 si el tiempo al evento de la unidad i es censurado y es 1 si el tiempo al evento es observado. La maximización de esta expresión respecto de los β's suministra los respectivos estimadores de máxima verosimilitud de los parámetros.

Para emplear el modelo (7.2.1) en el proceso de optimización, se deben conocer el valor de cada una de las covariables para todas las unidades en el conjunto de riesgo en cada tiempo, t_i, donde ocurre un evento.

Cuando la forma funcional de la variable j-ésima, $x_j(t)$, es conocida, resulta sencillo pronosticar su valor en un tiempo t_i. Cuando no se dispone explícitamente de la forma funcional particular de una variable, el valor que toma esta variable en un tiempo debe aproximarse. Hay varias formas de hacer esta aproximación.

Una salida es usar el último valor de la variable registrado antes del tiempo requerido. Cuando la variable ha sido registrada para una unidad antes y después del tiempo requerido, se toma el valor más cercano en el tiempo. Otra opción es usar la interpolación lineal entre valores consecutivos de la variable. Esta interpolación no es adecuada cuando la variable es de tipo categórico; en este caso, se puede asignar el valor con mayor probabilidad, el cual se puede calcular con regresión logística.

Modelo de Cox como proceso de conteo

En el modelo de riesgos proporcionales de Cox clásico se asume que la función de riesgo toma la forma

$$h_i(t|\mathbf{x}_i(t)) = h_0(t)e^{x_i'(t)\beta},$$

donde $h_0(t)$ es una función no negativa sin especificar.

La formulación del modelo de Cox como un *proceso de conteo* reemplaza el par de variables $(\mathbf{x}_i, \delta_i)$ por el par de funciones $(N_i(t), Y_i(t))$, donde

$$N_i(t) = \text{el número de eventos observados en } [0, t] \text{ para la unidad } i$$

$$Y_i(t) = \begin{cases} 1, & \text{la unidad } i \text{ está bajo observación y en riesgo al tiempo } t \\ 0, & \text{en otro caso.} \end{cases}$$

Esta formulación incluye los datos censurados a derecha como un caso especial; $N_i(t) = I(\{\mathbf{T}_i \leq t, \delta_i = 1\})$ y $Y_i(t) = I(\{\mathbf{T}_i \geq t\})$, con $I(\cdot)$ como la función indicadora.

La variable aleatoria $N(t)$ representa un *proceso de conteo* sobre $[0, \infty]$ si

1. $N(t)$ es un entero no-negativo, también se asume que $N(0) = 0$;
2. $N(s) \leq N(t)$ para $s < t$;
3. $dN(t) = N(t) - N(t^-)$ es 0 o 1, donde $N(t^-)$ denota a $\lim_{\delta \downarrow 0} N(t - \delta)$;
4. $\mathscr{E}(N(t)) < \infty$.

El proceso de conteo observado $N_i = \{N_i(t) : t \geq 0\}$ cuenta el número de eventos de la unidad i, la ocurrencia se asume dentro del intervalo $(0, t]$. Un proceso de conteo puede ser considerado como el proceso que registra y cuenta el número de eventos sobre una unidad en el tiempo de estudio.

En el modelamiento clásico de sobrevida, las principales variables incluidas son la variable tiempo para evento T_i^*; el tiempo de censura C_i, con $T_i = \text{mín}\{T_i^*, C_i\}$; y la variable indicadora δ_i, la cual es igual a 1 si T_i^* es observada y 0 si la observación es censurada. Así, los datos consisten en el par (T_i, δ_i). Ahora, con la formulación del proceso de conteo, el par (T_i, δ_i) es reemplazado por el par $(N_i(t), \rho_i(t))$, donde $N_i(t)$ es el número de eventos en el intervalo $[0, t]$ para la unidad i y $\rho_i(t)$ es un indicador de riesgo en el tiempo t.

Ejemplo 7.2.1. Se consideran los datos de un estudio de cirrosis tomados de Collet [10, pp. 266-267]. Doce pacientes se incluyen en un estudio sobre el tratamiento de cirrosis en el hígado. A los pacientes se les asignó aleatoriamente un placebo o un tratamiento nuevo referido como Liverol. Seis pacientes reciben aleatoriamente Liverol y los otros seis reciben un placebo. Al momento de ingresar los pacientes al estudio, la edad y el nivel de bilirrubina de los pacientes son registrados. El logaritmo natural del valor de la bilirrubina (en $\mu mol/l$) se usa en el análisis. Las variables son las siguientes:

- Tiempo de sobrevida del paciente en días.
- Indicador del evento ($\delta_i = 0$ censura y $\delta_i = 1$ no censura).
- Tratamiento (0 = placebo, 1 = Liverol).
- Edad del paciente en años.
- Logaritmo del nivel de bilirrubina.

Los datos asociados con los 12 pacientes y las variables están en la tabla 7.1, en la cual las covariables no son dependientes del tiempo.

En el estudio se espera que los pacientes regresen a la clínica a los tres, seis y doce meses después de iniciar el tratamiento, y luego anualmente. En estas visitas se mide el nivel de bilirrubina y se registra, de manera que la variable *bilirrubina* es una variable dependiente del tiempo, es decir, cambia durante el seguimiento del paciente. En la tabla 7.2 se muestran los valores del logaritmo de la bilirrubina durante el seguimiento para cada uno de los 12 pacientes.

El logaritmo de la bilirrubina (*Lbrt*) es una variable dependiente del tiempo, el valor de la variable en cualquier tiempo t es el valor registrado en la última visita de seguimiento inmediatamente antes del tiempo t, para cada paciente. El cambio de *Lbrt* en un paciente a un valor nuevo se asume que toma lugar inmediatamente después de que la medición ha sido realizada sobre el paciente. Por ejemplo, cuando el paciente 1 inicia el estudio tiene un $log(bilirrubina)$ de 3.2. En el día 47, después de iniciar el estudio, el valor es de 3.8, en el día 184 el valor de $log(bilirrubina)$ es 4.9 y, finalmente, en el día 251 esta variable toma el valor de 5.0. Esto es equivalente a suponer que los valores de *Lbr* para un paciente siguen una función escalonada en la cual los valores se asumen constantes entre cualquier par de puntos de medición consecutivos.

En la figura 7.1 se muestra la "trayectoria" de la variable *Lbrt* para el paciente núm. 1 y se observa que es una función escalonada continua por la derecha. Los datos se han estructurado teniendo en cuenta esta característica, la cual se emplea en el paquete R para la estimación de los modelos de Cox con variable dependiente del tiempo. La tabla 7.8, que se muestra al final del capítulo, contiene los datos así estructurados.

En primera instancia, de manera independiente, se ajustan los modelos de la función de riesgo con las siguientes covariables: (i) sin covariables (modelo nulo), (ii) *edad*, (iii) *logaritmo de la bilirrubina* (*Lbr*), (iv) *edad* y *logaritmo de la bilirrubina* y (v) *Edad, logaritmo de la bilirrubina* (*Lbr*) y *tratamiento* (*trto*). Se nota que todas las covariables no son dependientes del tiempo. Para cada

modelo se calcula la estadística $-2\ln\hat{L}$, los resultados se muestran en la tabla 7.3. A partir de esta tabla se concluye que la función de riesgo depende moderadamente de las covariables *edad* y *logaritmo de la bilirrubina* (*Lbr*). No obstante, la inclusión simultánea de estas covariables en un mismo modelo no es muy significativa. Cuando se agrega la covariable *tratamiento* (*trto*) al modelo que contiene *edad* y *Lbr*, la reducción en el valor de la estadística $-2\ln\hat{L}$ se reduce en $18.47542 - 13.29283 = 5.18259$. Esta es una reducción significativa al 5 %, pues el respectivo valor p es igual a 0.02281434. El coeficiente del tratamiento es -3.0519 y significativo (valor p igual a 0.051), lo cual indica que el medicamento Liverol es efectivo en la reducción del riesgo al evento muerte. Así, Liverol reduce el riesgo a la muerte por un factor de $e^{-3.019} = 0.0473$. Se puede decir que este medicamento, a partir de estos datos, tiene un efecto "protector" sobre la muerte.

Tabla 7.1. Tiempo de sobrevida de pacientes con cirrosis de hígado

Paciente	Tiempo	Censura	Tratamiento	Edad	Lbr
1	281	1	0	46	3.2
2	604	0	0	57	3.1
3	457	1	0	56	2.2
4	384	1	0	65	3.9
5	341	0	0	73	2.8
6	842	1	0	64	2.4
7	1514	1	1	69	2.4
8	182	0	1	62	2.4
9	1121	1	1	71	2.5
10	1411	0	1	69	2.3
11	814	1	1	77	3.8
12	1071	1	1	58	3.1

Tomado de Collet [10, p. 267]

El código R empleado para hacer el procesamiento estadístico es:

```
## MODELO DE COX CON DATOS SIN VARIABLES TPO DEPENDIENTES
cirro0<-read.table("cirrosis0.txt",header=T) # Leer Tabla 7.1
attach(cirro0)
require(survival)
```

Tabla 7.2. Seguimiento de la bilirrubina en pacientes con cirrosis de hígado

Paciente	Tiempo	log(bilirrubina)	Paciente	Tiempo	log(bilirrubina)
1	47	3.8	7	74	2.9
	184	4.9		202	3.0
	251	5.0		346	3.0
2	94	2.9		917	3.9
	187	3.1		1411	5.1
	321	3.2	8	90	2.5
3	61	2.8		182	2.9
	97	2.9	9	101	2.5
	142	3.2		410	2.7
	359	3.4		774	2.8
	440	3.8		1043	3.4
4	92	4.7	10	182	2.2
	194	4.9		847	2.8
	372	5.4		1051	3.3
5	87	2.6		1347	4.9
	192	2.9	11	167	3.9
	341	3.4		498	4.3
6	94	2.3	12	108	2.8
	197	2.8		187	3.4
	384	3.5		362	3.9
	795	3.9		694	3.8

```
## MODELO DE COX SIN VARIABLES TIEMPO DEPENDIENTES
cox0<-coxph(Surv(tpo,Censura)~1)
cox0
cox0$loglik
summary(cox0)
cox1<-coxph(Surv(tpo,Censura)~edad)
cox1
cox1$loglik
summary(cox1)
cox2<-coxph(Surv(tpo,Censura)~Lbr)
cox2
cox2$loglik
```

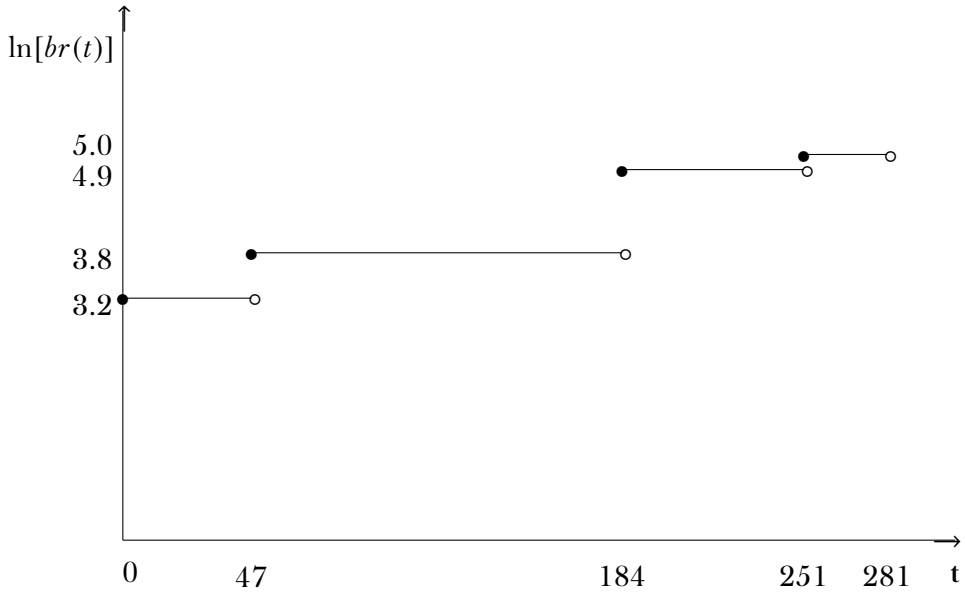

Figura 7.1. Log-bilirrubina (Lbrt) del paciente núm. 1.

Tabla 7.3. Modelos sin variable tiempo-dependiente

Modelo	Coefic.	$-2\ln\hat{L}$	ee(Coefic.)	Z	Valor p
Modelo Nulo		25.12134			
Edad	-0.1082	22.13538	0.0652	-1.66	0.0969
Lbr	1.43	21.66236	0.79	1.81	0.07
Edad y	-0.0999		0.0590	-1.69	0.09
Lbr	1.7701	18.47542	0.9753	1.81	0.07
Edad	-0.0850		0.0773	-1.10	0.271
Lbr y	2.5608		1.2721	2.01	0.044
Trto	-3.0519	13.29283	1.5615	-1.95	0.051

```
summary(cox2)
cox3<-coxph(Surv(tpo,Censura)~edad+Lbr)
cox3
cox3$loglik
summary(cox3)
cox4<-coxph(Surv(tpo,Censura)~Lbr+factor(trto))
cox4
cox4$loglik
summary(cox4)
```

La salida para el modelo que contiene las covariables *edad*, *Lbr* y *trto* es la siguiente:

```
coef exp(coef) se(coef)     z     p
edad          -0.0850    0.9185   0.0773 -1.10 0.271
Lbr            2.5608   12.9462   1.2721  2.01 0.044
factor(trto)1 -3.0519    0.0473   1.5615 -1.95 0.051
```

Tabla 7.4. Modelos con variable tiempo-dependiente

Modelo	Estadística $-2\ln\hat{L}$
Modelo nulo	25.12134
Edad	22.13538
Lbrt	12.04975
Edad y Lbrt	11.14492
Lbrt y Trto	10.6762

Ahora se desarrolla el modelamiento con la covariable tiempo-dependiente *logaritmo de la bilirrubina* (*Lbrt*) en el tiempo (tablas 7.1 y 7.2). Los valores de la estadística $-2\ln\hat{L}$ para cada uno de los modelos de Cox ajustados se presentan en la tabla 7.4.

El efecto de la variable *Lbrt* es evidente, pues el valor de la estadística $-2\ln\hat{L}$ se reduce a 22.13538 con la variable *edad*, y a 12.04975 con la variable *Lbrt*. Cuando la variable *edad* se incluye junto con la variable *Lbrt* su efecto es pequeño, puesto que $-2\ln\hat{L}$ pasa de 12.04975 a 11.14492. En consecuencia, se adiciona la variable *tratamiento* (*trto*) al modelo que contiene solo la variable *Lbrt*. El efecto de esta última adición es que $-2\ln\hat{L}$ se reduce de 12.04975 a 10.6762. La reducción es en 1.37355, la cual no

es significativa (valor p es igual a 0.2412029). Se concluye que la función de riesgo depende de la dinámica del logaritmo de la bilirrubina y ningún efecto del tratamiento es apreciable.

La función de riesgo para el i-ésimo paciente es

$$\hat{h}_i(t|Lbrt, trto) = \exp\{3.605Lbrt - 1.479trto\}h_0(t),$$

donde $Lbrt$, aunque debiera notarse como $Lbr(t)$, resalta la dependencia del nivel de bilirrubina en el tiempo t. La razón de riesgo a la muerte en el tiempo t, entre dos pacientes que reciben el mismo tratamiento y quienes tienen una diferencia de una unidad al momento t en el nivel de Lbr, es igual a $e^3.605 = 36.773$. Una explicación de la diferencia entre los modelos con y sin variable dependiente del tiempo (tablas 7.3 y 7.4, respectivamente) es que el efecto del tratamiento es cambiar los valores del nivel de bilirrubina, de forma tal que, después de los cambios en estos valores a través del tiempo, no se evidencie el efecto del tratamiento.

La tabla de datos 7.8 ha sido estructurada en la forma "start/stop" (en la tabla ti y tf) para permitir su procesamiento con el paquete R. El código R para el desarrollo de los respectivos cómputos de los modelos, cuyos resultados son editados en la tabla 7.4, se muestra a continuación.

```
## DATOS DE CIRROSIS CON VARIABLES TPO DEPENDIENTES
cirro1<-read.table("Tabla 7.5",header=T)
attach(cirro1)
require(survival)
## MODELOS DE COX CON VARIABLE TIEMPO DEPENDIENTE
coxt0<-coxph(Surv(time=ti,time2=tf,cen)~1,
+method="breslow")
coxt0
coxt0$loglik
coxt1<-coxph(Surv(time=ti,time2=tf,cen)~Edad,
+method="breslow")
coxt1
coxt1$loglik
coxt2<-coxph(Surv(time=ti,time2=tf,cen)~Lbrt,
+method="breslow")
coxt2
coxt2$loglik
coxt3<-coxph(Surv(time=ti,time2=tf,cen)~Edad+Lbrt,
+method="breslow")
coxt3
```

```
coxt3$loglik
coxt4<-coxph(Surv(time=ti,time2=tf,cen)~Lbrt+factor(Trto),
+method="breslow")
coxt4
coxt4$logli
```

La función de riesgo base acumulada, $H_0(t)$, puede estimarse para el modelo que contiene la covariable dependiente del tiempo, el logaritmo de la bilirrubina, *Lbrt*, y la covariable *tratamiento*, *trto*. La tabla 7.5 contiene los valores de esta función. Se observa que el riesgo base no es constante sino que se incrementa en el tiempo.

La figura 7.3 contiene las funciones de riesgo acumulado base estimadas para los grupos "tratamiento" (Liverol) y "placebo", respectivamente. Estas funciones se evalúan y grafican para un paciente cuyo valor del logaritmo de la bilirrubina es $Lbr = 3.0$, en los dos grupos.

Tabla 7.5. Estimación de $H_0(t)$ en estudio de cirrosis

Tiempo de seguimiento	$\widehat{H}_0(t)$
0	0.000
281	8.716865×10^{-9}
384	1.222210×10^{-8}
457	5.413506×10^{-7}
814	9.084826×10^{-7}
842	1.577272×10^{-6}
1071	3.318074×10^{-6}
1121	6.007338×10^{-6}
1514	6.052855×10^{-6}

Las correspondientes funciones de sobrevida también se evaluaron y graficaron para una persona con $Lbr = 3.0$. Estas funciones se muestran en la figura 7.3. Allí se muestra que el grupo de pacientes que recibieron el placebo tienen un pronóstico más bajo, en términos de probabilidad de sobrevida, que los del grupo que recibe el tratamiento con Liverol.

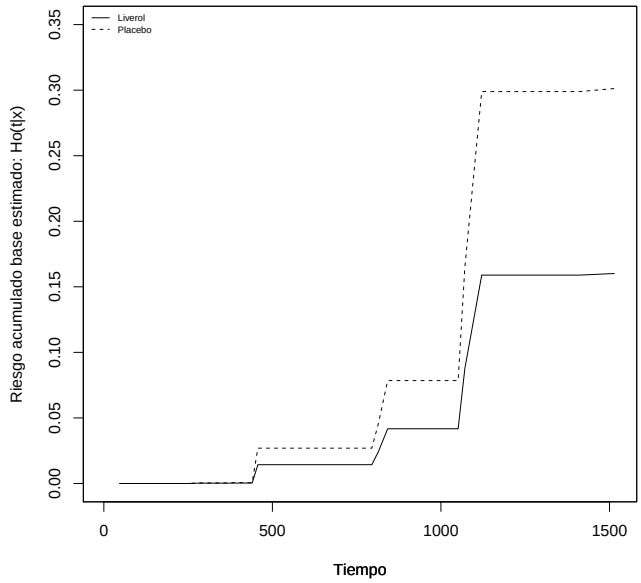

Figura 7.2. Funciones de riesgo acumulado base estimadas por tratamiento de cirrosis.

7.3. Modelos estratificados

El supuesto estructural del modelo de Cox es que la razón de las funciones de riesgo en cualquier par de unidades i e i' con covariables de pronóstico $\mathbf{x}_i$ e $\mathbf{x}_i'$ es una constante, independiente del tiempo. Este supuesto no siempre es satisfecho por los datos asociados con situaciones prácticas. Para acomodarse al caso de no proporcionalidad, el modelo de Cox puede extenderse empleando el concepto de *estratificación* mostrado por Kalbfleisch y Prentice [37, p. 118].

Los estratos pueden tenerse de entrada o no, de acuerdo con la naturaleza de las covariables; así es para las variables con naturaleza categórica, tales como un tratamiento con diferentes niveles, el sexo, el tipo de actividad laboral, la raza, entre otros. En el caso de las covariables de naturaleza cuantitativa (al menos en escala de intervalo), estas se pueden categorizar mediante la introducción de puntos de corte que determinen las categorías o estratos, por ejemplo, la variable estatura se puede categorizar en los valores "baja", "media" y "alta". Lo mismo puede decirse para variables como el peso, los ingresos o los niveles de colesterol de las personas; o la dureza

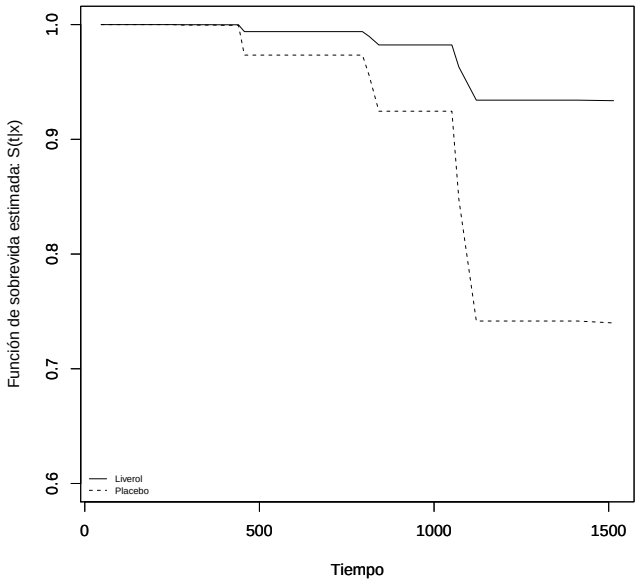

Figura 7.3. Funciones de sobrevida estimadas por tratamiento de cirrosis.

y la conductividad eléctrica en materiales; o la humedad, la temperatura, la concentración de gas carbónico y el brillo solar en covariables ambientales.

Se advierte que los estratos no solo se generan por la naturaleza de las covariables sino por el análisis del supuesto de proporcionalidad del modelo, orientado e indicado por el análisis de residuales mostrado en la sección 6.4.

En resumen, los estratos dividen las unidades en grupos disjuntos cada uno de los cuales tiene una función de riesgo base diferente pero valores comunes para los coeficientes del vector β. Los estratos juegan un papel similar al de los bloques en un diseño de bloques aleatorizado dentro del análisis de varianza a dos vías de clasificación en el diseño experimental.

La estrategia de modelar por estratos consiste en que los modelos de Cox sean de riesgos proporcionales dentro de cada estrato. Por ejemplo, la proporcionalidad del modelo se tiene tanto dentro del estrato conformado por mujeres como el constituido por hombres.

Un modelamiento estratificado consiste en dividir los datos de tiempo para el evento en m estratos, de acuerdo con la indicación del incumplimiento del supuesto de proporcionalidad. El modelo de riesgos proporcionales

de Cox *estratificado* se expresa entonces como

$$h_{ij}(t|\mathbf{x}) = h_{0j}(t)\exp\{\mathbf{x}'_{ij}\beta\},\ j = 1\ldots, m, i = 1, \ldots, n_j, \tag{7.3.1}$$

donde n_j es el número de unidades en el estrato $j = 1, \ldots, m$.

Las funciones de riesgo base $h_j(t)$ se asumen como diferentes para cada estrato y no relacionadas. No obstante, los coeficientes de regresión de los β's son los mismos en todos los estratos. Es decir, se asume que los riesgos son proporcionales para unidades dentro de cada estrato, pero no entre estratos o niveles.

Computacionalmente, la función de verosimilitud parcial de todas las observaciones desde los m estratos es definida como

$$L(\beta) = \prod_{j=1}^{m} L_j(\beta), \tag{7.3.2}$$

donde $L_j(\beta)$ es la verosimilitud parcial o marginal para el j-ésimo estrato. El logaritmo de la función de verosimilitud es, entonces,

$$l(\beta) = \ln L(\beta) = \sum_{j=1}^{m} l_j(\beta), \tag{7.3.3}$$

donde $l_j(\beta) = \ln L_j(\beta)$ se obtiene empleando solamente los datos de las unidades que pertenecen al estrato $j = 1, \ldots, m$. Las estimaciones de β se obtienen como solución del sistema de ecuaciones de derivadas parciales igualadas a cero, $\partial l(\beta)/\partial\beta_j = 0$. Una aproximación a la solución de estas ecuaciones se obtiene mediante el método de Newton-Raphson.

El vector de puntuación y la matriz de información son sumas similares:

$$U(\beta) = \sum_{j=1}^{m} U_j(\beta) \text{ e } \mathcal{I}(\beta) = \sum_{j=1}^{m} \mathcal{I}_j,$$

respectivamente.

En el modelo de Cox estratificado (7.3.1) se asume que las covariables tienen un efecto similar en la función de riesgo base de cada estrato, de ahí que los β's se asumen igual para todos los estratos. Este supuesto puede ser verificado mediante el empleo de la razón de verosimilitudes, a través de la estadística

$$RV = -2\left[l(\hat{\beta}) - \sum_{i=1}^{m} l_j(\hat{\beta}_j)\right], \tag{7.3.4}$$

donde $l(\hat{\beta})$ es el logaritmo de la función de verosimilitud parcial sobre el modelo que asumen los β's comunes en cada estrato y $\sum_{i=1}^{m} l_j(\hat{\beta}_j)$, es el logaritmo de la función de verosimilitud parcial sobre el modelo que asumen β's diferentes en cada estrato.

Ejemplo 7.3.1. Considere los tiempos y datos de sobrevida (tiempo hasta la muerte) de pacientes con leucemia mieloide aguda (en inglés AML), tomados del libro de Lee y Wang [44, p. 275] y consignados en la tabla 7.6. Dos posibles factores de pronóstico o covariables, *edad* y *celularidad*, se han definido de la siguiente forma:

$$X_1 = \begin{cases} 1, & \text{si } edad \geq 50 \\ 0, & \text{en otro caso} \end{cases}, \quad X_2 = \begin{cases} 1, & \text{si celularidad de médula es } 100\,\% \\ 0, & \text{en otro caso.} \end{cases}$$

Tabla 7.6. Tiempos de sobrevida de pacientes con AML

Tiempo	Cens	X_1	X_2	Tiempo	Cens	X_1	X_2
18	1	0	0	8	1	1	0
9	1	0	1	2	1	1	1
28	0	0	0	26	0	1	0
31	1	0	1	10	1	1	1
39	0	0	1	4	1	1	0
19	0	0	1	3	1	1	0
45	0	0	1	4	1	1	0
6	1	0	1	18	1	1	1
8	1	0	1	8	1	1	1
15	1	0	1	3	1	1	1
23	1	0	0	14	1	1	1
28	0	0	0	3	1	1	0
7	1	0	1	13	1	1	1
12	1	1	0	13	1	1	1
9	1	1	0	35	0	1	0

Suponga que no se tiene seguridad sobre si el riesgo a la muerte de pacientes cuya edad es al menos 50 años es proporcional a la de los pacientes con edad menor a 50 años; por tanto, se decide hacer una estratificación

respecto a la variable *edad*. Las funciones de riesgo se asumen como:

$$h_1(t|X_2) = h_{01}(t)\exp\{\beta_2 X_2\},$$
$$h_2(t|X_2) = h_{02}(t)\exp\{\beta_2 X_2\},$$

donde $h_1(t|X_2)$ es la función de riesgo para pacientes cuya edad es menor a 50 años y $h_2(t|X_2)$ es la función de riesgo para pacientes cuya edad es mayor o igual a 50 años; ambas funciones dependen de la celularidad, X_2. Las funciones $h_{01}(t)$ y $h_{02}(t)$ son las funciones de riesgo base para cada uno de los dos grupos etáreos.

El código R para hacer los cálculos de los modelos de Cox estratificados por la variable edad (X_1) es:

```
aml<-read.table("Tabla 7.6.txt",header=T)
#Tabla 7.6 se debe estructurar para leer en R
attach(aml)
require(survival)
# MODELO ESTRATIFICADO POR LA VARIABLE EDAD (X1)
cox_est2<-coxph(Surv(Tpo, Cens)~factor(X2)+strata(X1),
+method="breslow")
summary(cox_est2)
```

Los resultados del procesamiento en R, con los dos estratos definidos por la edad X_1, son:

```
              coef exp(coef) se(coef)     z Pr(>|z|)
factor(X2)1 0.2173    1.2427   0.4411 0.493    0.622
```

Estos resultados no difieren de los obtenidos para el modelo sin estratificar, los cuales se muestran en la siguiente salida de R:

```
# RESULTADOS DEL MOLDELO SIN ESTRATIFICAR
              coef exp(coef) se(coef)     z Pr(>|z|)
factor(X1)1 1.0132    2.7544   0.4574 2.215   0.0268 *
factor(X2)1 0.3503    1.4194   0.4392 0.798   0.4251
```

7.4. Modelo de Cox con censura por intervalo

Hasta ahora los datos que se han modelado han estado sujetos solo a censura por la derecha. Es decir, cada observación sobre la unidad (persona,

animal, planta u objeto) en el tiempo comienza en un momento bien definido y conocido (origen), y el seguimiento continua hasta que el evento de interés ocurre o hasta que la observación sobre la unidad termina por motivos no relacionados con el evento de interés. Por ejemplo, el sujeto se retira o el estudio finaliza. En la práctica, este es el tipo de censura más frecuentemente encontrado en los datos de tiempo hasta el evento; sin embargo, hay otras razones por las que se tienen otras clases de observaciones incompletas del tiempo al evento. En esta sección, consideramos algunas de estas observaciones, así como los métodos que extienden el modelo de riesgos proporcionales para abordarlos.

Suponga que la escala de tiempo para un estudio se representa como una línea horizontal. Una observación incompleta del tiempo al evento puede ocurrir en cualquier punto de la línea, aunque es más común al principio (es decir, a la izquierda) o al final (es decir, a la derecha) de la escala de tiempo. La censura y el truncamiento son las dos causas más comunes de observaciones incompletas. Así, cuando se considera una observación del tiempo al evento, se debe evaluar si la observación está sujeta a censura o truncamiento a la izquierda, como también a censura o truncamiento a la derecha. En la práctica, es poco frecuente que una observación sea censurada y truncada en el mismo punto de la escala de tiempo; sin embargo, un evento puede estar sujeto a observación incompleta en ambos lados (por ejemplo, truncado a izquierda y censurado a derecha).

Se asume que el tiempo para el evento de la i-ésima unidad, $\mathbf{T}_i$, está acotado entre dos valores conocidos, denotado por $a_i T@ < \mathbf{T_i} < \mathbf{b_i}$. Además, se conoce si el evento de interés ha ocurrido, para esto se emplea la función indicadora de censura δ_i, para la cual $\delta_i = 1$, si ocurre el evento de interés, y $\delta_i = 0$, en el caso contrario. Para observaciones que son censuradas a izquierda se tiene $a_i = 0$ y $\delta_i = 1$. En las observaciones con censura a derecha, $b_i = \infty$ y $\delta_i = 0$.

La función de sobrevida al tiempo t para la unidad identificada con el vector de covariables $\mathbf{x}_i$, con vector de parámetros β, se escribe como $S(t, \mathbf{x}_i)$. La probabilidad en el intervalo observado para la i-ésima unidad es

$$S(a_i, \mathbf{x}_i, \beta)^{1-\delta_i}[S(a_i, \mathbf{x}_i, \beta) - S(b_i, \mathbf{x}_i, \beta)]^{\delta_i}. \qquad (7.4.1)$$

A partir de (7.4.1), se obtiene $[1 - S(b_i, \mathbf{x}_i, \beta)]$ para datos con censura a izquierda; ($a_i = 0$, $\delta_i = 1$ y $S(0, \mathbf{x}_i, \beta) = 1$) para observaciones censuradas por la derecha; $S(a_i, \mathbf{x}_i, \beta)$, ($\delta_i = 0$) y $[S(a_i \mathbf{x}_i, \beta) - S(b_i, \mathbf{x}_i, \beta)]$ para observaciones no censuradas. No obstante, esta última expresión puede corresponder al caso de censura por intervalo, cuando a_i y b_i están dentro del periodo de estudio, $[t_0, t_f]$, y $\delta_i = 1$.

El primer paso para obtener estimadores de los parámetros en un modelo de regresión es construir la función de verosimilitud y luego evaluarla con el modelo escogido. Se usará el modelo de riesgos proporcionales, pero el método es lo suficientemente general para permitir emplear otros modelos, incluso los modelos paramétricos presentados en el capítulo 5. La verosimilitud es el producto de los términos obtenidos al evaluar (7.4.1) en cada unidad, así:

$$L(\beta) = \prod_{i=1}^{n} S(a_i, \mathbf{x}_i, \beta)^{1-\delta_i} [S(a_i, \mathbf{x}_i, \beta) - S(b_i, \mathbf{x}_i, \beta)]^{\delta_i}. \qquad (7.4.2)$$

Los cálculos y el ajuste del modelo se simplifican si tan solo se toman unos pocos valores para los a y los b; así, es más fácil referirse a los valores de los intervalos de censura en la escala de tiempo común para todas las unidades. Se asume que se tienen $J + 1$ de tales intervalos, denotados por $(t_{j-1}, t_j]$ para $j = 1, \ldots, J + 1$, con $t_0 = 0$ y $t_{J+1} = \infty$. Además, estos intervalos son los mismos para todas las unidades.

Para simplificar, se considera que I_j representa el intervalo $(t_{j-1}, t_j]$. La variable dicotómica y_{ij} indica si el intervalo de tiempo observado para la i-ésima unidad es I_j. Se define por

$$y_{ij} = \begin{cases} 1, & \text{si } (a_i, b_i] = I_j \\ 0, & \text{en otro caso.} \end{cases}$$

Mediante una factorización apropiada se expresa la probabilidad para el intervalo I_j como

$$[S(t_{j-1}, \mathbf{x}_i, \beta) - S(t_j, \mathbf{x}_i, \beta)] = S(t_{j-1}, \mathbf{x}_i, \beta) \left[1 - \frac{S(t_j, \mathbf{x}_i, \beta)}{S(t_{j-1}, \mathbf{x}_i, \beta)}\right]. \qquad (7.4.3)$$

El segundo término dentro del paréntesis rectangular en (7.4.3) es la probabilidad condicional de que el evento ocurra en el j-ésimo intervalo, dado que la unidad no ha tenido el evento al extremo derecho del intervalo $j - 1$, es decir,

$$P(t_{j-1}T@ < \mathbf{T} \leq \mathbf{t_j} | \mathbf{T} > \mathbf{t_{j-1}}).$$

Bajo el modelo de riesgos proporcionales, de acuerdo con la ecuación (6.2.2), la función de sobrevida es $S(t|\mathbf{x}) = [S_0(t)]^{\exp(\mathbf{x}'\beta)}$. La razón de las funciones de sobrevida de intervalos con extremos (después de las transformaciones algebraicas) se puede escribir como

$$\frac{S(t_j, \mathbf{x}_i, \beta)}{S(t_{j-1}, \mathbf{x}_i, \beta)} = \exp\{-\exp(\mathbf{x}_i'\beta + \tau_j)\}, \qquad (7.4.4)$$

donde

$$\tau_j = \ln\left(-\ln\left[\frac{S_0(t_j)}{S_0(t_{j-1})}\right]\right).$$

Para simplificar la notación, se escribe la probabilidad condicional de (7.4.3) como

$$\theta_{ij} = 1 - \exp\{-\exp(\mathbf{x}_i'\beta + \tau_j)\}. \tag{7.4.5}$$

De la expresión (7.4.3), la función de verosimilitud (7.4.2) se puede escribir en la forma

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n}\prod_{j=1}^{J+1}\left\{\left[S(t_{j-1}, \mathbf{x}_i, \beta)^{1-\delta_i}\right]\left[S(t_{j-1}, \mathbf{x}_i, \beta) - S(t_j, \mathbf{x}_i, \beta)\right]^{\delta_i}\right\}^{y_{ij}} \\ &= \prod_{i=1}^{n}\prod_{j=1}^{J+1}\left\{\left[S(t_{j-1}, \mathbf{x}_i, \beta)\right]\left[\left[1 - \frac{S(t_j, \mathbf{x}_i, \beta)}{S(t_{j-1}, \mathbf{x}_i, \beta)}\right]\right]^{\delta_i}\right\}^{y_{ij}} \\ &= \prod_{i=1}^{n}\prod_{j=1}^{J+1}\left\{S(t_{j-1}, \mathbf{x}_i, \beta) \times \theta_{ij}^{\delta_i}\right\}^{y_{ij}}. \end{aligned} \tag{7.4.6}$$

Mediante un procedimiento similar al seguido para el estimador de Kaplan-Meier, como se muestra en la sección 2.3, la función de sobrevida en un punto del extremo de un intervalo se expresa como un producto de probabilidades de sobrevida condicionada en intervalos consecutivos, empleando la expresión (7.4.5). Después de algunos detalles algebraicos, el resultado es el siguiente:

$$S(t_{j-1}, \mathbf{x}_i, \beta) = \prod_{l=1}^{j-1}(1 - \theta_{il}). \tag{7.4.7}$$

Se reemplaza la expresión contenida en (7.4.7) en la función de verosimilitud (7.4.6) para obtener la siguiente función de verosimilitud:

$$L(\beta) = \prod_{i=1}^{n}\prod_{j=1}^{J+1}\left[\prod_{l}^{j-1}(1 - \theta_{il})\theta_{il}^{\delta_i}\right]^{y_{ij}}. \tag{7.4.8}$$

Sea el intervalo observado para la i-ésima unidad denotado por I_{k_i}, es decir, $I_{k_i} = (a_i, b_i]$.

Lo primero que se debe tener en cuenta en (7.4.8) es que la única vez que los términos en el producto sobre j difieren de l es cuando $j = k_i$; en este caso, $y_{ij} = 1$ y $y_{ij} = 0$ en los demás términos. Luego, la expresión de verosimilitud (7.4.8) se reduce a

$$L(\beta) = \prod_{i=1}^{n} \theta_{ik_i}^{\delta_i} \prod_{j=1}^{k_i-1} (1 - \theta_{ij}). \tag{7.4.9}$$

La función de verosimilitud (7.4.9) puede asimilarse a la función de verosimilitud para un modelo de regresión binaria. Esto se logra mediante la definición de una variable respuesta "cuasi-binaria" como $z_{ij} = y_{ij} \times \delta_i$ y se emplea en (7.4.9) para obtener

$$L(\beta) = \prod_{i=1}^{n} \prod_{j=1}^{k_i-1-\delta_i} (1 - \theta_{ij})^{1-z_{ij}} \theta_{ik_i}^{z_{ij}}. \tag{7.4.10}$$

Para cada unidad i, la expresión (7.4.10) es la verosimilitud de las $k_i - 1 + \delta_i$ observaciones binarias independientes con probabilidades θ_{ij} y respuestas z_{ij}.

De esta manera, se pueden emplear las herramientas de cómputo usuales (como R y SAS) para ajustar un modelo de regresión de riesgos proporcionales (modelo de Cox) en datos con censura por intervalo.

Ejemplo 7.4.1. Se trata de un estudio retrospectivo llevado a cabo para comparar los efectos cosméticos de la radioterapia sola frente a la radioterapia combinada con quimioterapia en mujeres con cáncer de mama temprano [39, p. 18]. El uso de una biopsia por escisión, seguida de radioterapia, ha sido sugerido como una alternativa a la mastectomía. Esta terapia preserva la mama y, por lo tanto, tiene el beneficio de un efecto cosmético mejorado. El uso de quimioterapia adyuvante a menudo está indicado para prevenir la recurrencia del cáncer, pero existe alguna evidencia clínica de que mejora los efectos de la radiación en el tejido normal, por lo tanto, compensa el beneficio cosmético de este procedimiento.

Para comparar los dos tratamientos, se hizo un estudio retrospectivo de 46 pacientes que recibieron solo radiación y 48 pacientes que recibieron radiación más quimioterapia. Las pacientes fueron observadas inicialmente cada 4 a 6 meses, pero a medida que la recuperación progresó, el intervalo entre visitas se alargó. En cada visita se registró una medida del deterioro mamario en una escala de tres puntos ("ninguno"= 1, "moderado"= 2 y "severo"= 3). El evento de interés fue el momento de una primera aparición de retracción mamaria moderada o grave.

Debido al hecho de que las pacientes fueron observadas solo en estos momentos aleatorios, se sabe que la retracción (disminución del volumen) del seno ocurre en el intervalo entre visitas. Así, este tipo de datos son censurados por intervalo.

Los datos consisten en el intervalo, en meses, en el cual se produjo el deterioro o la última vez que se ha visto a una paciente sin que se haya producido ningún deterioro (censura a derecha, se nota como "≥"). Los datos para los dos grupos se muestran a continuación.

Radioterapia únicamente: (0, 7]; (0, 8]; (0, 5]; (4, 11]; (5, 12]; (5, 11]; (6, 10]; (7, 16]; (7, 14]; (11, 15]; (11, 18]; ≥15; ≥17; (17,25]; (17, 25]; ≥18; (19,35]; (18, 26]; ≥22; ≥24; ≥24; (25,37]; (26,40]; (27, 34]; ≥32; ≥33; ≥34; (36,44]; (36, 48]; ≥36; ≥36; (37, 44]; ≥37; ≥37; ≥37; ≥38; ≥40; ≥45; ≥46; ≥46; ≥46; ≥46; ≥46; ≥46; ≥46; ≥46.

Radioterapia + quimioterapia: (0, 22]; (0, 5]; (4,9]; (4,8]; (5,8]; (8, 12]; (8,21]; (10,35]; (10,17]; (11,13]; ≥11; (11,17]; ≥11; (11,20]; (12,20]; ≥13; (13,39]; ≥13; ≥13; (14,17]; (14, 19]; (15,22]; (16,24]; (16,20]; (16,24]; (16,60]; (17,27]; (17,23]; (17,26]; (18,25]; (18,24]; (19, 32]; ≥21; (22, 32]; ≥23; (24,31]; (24,30];; (30, 34];; (30, 36];; ≥31; ≥32; (33,40];; ≥34; ≥34; ≥35; (35,39]; (44, 48]; ≥48.

Considerando la covariable *tratamiento* ($terap$) que asume el valor 1 si la paciente recibe radioterapia y asume el valor 0 si la paciente recibe radioterapia y quimioterapia, se ajusta el modelo de Cox escrito como

$$h(t|terap) = h_0(t)\exp\{\beta_1 terap\}.$$

El paquete "icenReg" de R (acrónimo de *regresión con datos censurados por intervalo*) es un procedimiento computacional para ajustar modelos de regresión de datos de tiempo hasta un evento con censura por intervalo.

El código R para obtener la estimación de los β's desde los datos anteriores es:

```
#Estructurar Tabla 7.7 para leerse en R
cancer<-read.table("Tabla 7.8.txt",header=T)
attach(cancer)
# Paquete de regresi\'on para datos con censura por intervalos
require(icenReg)
coxp_int<- ic_sp(Surv(izq, der, type ="interval2") ~ terap,
data =cancer, B=c(0,1),bs_samples = 1000)
```

Los resultados obtenidos mediante el procedimiento "icenReg" son:

```
     Estimate Exp(Est) Std.Error z-value       p
terap  -0.7949   0.4516    0.3352  -2.372 0.01771
```

Un intervalo de confianza del 95 % para β_1, empleando el error estándar obtenido mediante las 1000 remuestras (*bootstrap*) es (−1.4421; −0.1477).

7.5. Modelos de fragilidad o vulnerabilidad

Se dice que una unidad es *frágil* si es más susceptible (es decir, está más expuesta al riesgo) de que le ocurra un determinado evento en comparación con otras unidades.

Vaupel, Manton y Stallard [69] introdujeron el término *fragilidad* para indicar que diferentes unidades están en riesgo a pesar de que, en apariencia, pueden parecer bastante similares con respecto a variables o atributos como edad, sexo, peso, etc. Ellos utilizaron el término *fragilidad* para representar un efecto aleatorio no observable (no medido) compartido por unidades con riesgos similares en el análisis de las tasas de ocurrencia del evento. Los modelos de fragilidad han sido introducidos en la literatura estadística con el propósito de dar cuenta sobre la posible existencia de atributos no medidos que introducen heterogeneidad en una población de estudio. Por lo tanto, la fragilidad o los efectos aleatorios de los modelos intentan dar cuenta de las correlaciones dentro de un grupo o unos grupos.

La comparación de historias de eventos y tiempos de sobrevida entre los miembros de una población puede advertir acerca de la heterogeneidad entre ellos, en lo que concierne a los factores de riesgo relevantes. La última fuente de variabilidad es conocida como vulnerabilidad, propensión, susceptibilidad o fragilidad (en inglés, *frailty*).

Por ejemplo, en estudios clínicos con un evento como la muerte o la recurrencia de una dolencia, algunos pacientes sobrevivirán o permanecerán saludables durante un tiempo largo a pesar de los factores de riesgo observables a los que puedan estar expuestos, mientras que algunos otros sobrevivirán tiempos más cortos de lo esperado. Esta heterogeneidad puede alterar la estimación del riesgo o la tasa del evento de interés, lo cual no representa el riesgo real de cualquier individuo particular o de un subgrupo de la población.

Una característica de esto es que las unidades tienen diferente grado de vulnerabilidad o de fragilidad, y que aquellas cuya vulnerabilidad sea mayor alcanzarán el evento de interés más pronto que las demás. Los modelos de vulnerabilidad se emplean también para hacer ajustes de sobredispersión[1] en los datos de sobrevivencia.

Se puede atribuir la fragilidad o vulnerabilidad a un conjunto de covariables explícitas o latentes, tales como las variables biológicas, genéticas o ambientales en los seres vivos; las características físicas o químicas en cier-

[1] Es decir, cuando las observaciones exhiben una variabilidad que supera a la pronosticada por el modelo [2, pp. 126-127].

tos materiales; las variables sociales o de comportamiento en personas o animales. No obstante, siempre se encontrará que la variabilidad, y por ende la diferencia en tiempo a un evento, se atribuye a causas no observables o consideradas. Se requiere, entonces, modelar esta variabilidad, y los *modelos de fragilidad* son una alternativa.

En la aplicación de los métodos estándar de análisis de sobrevivencia, principalmente en la investigación clínica, se asume que la población estudiada es homogénea aún cuando los sujetos estén sometidos a diferentes condiciones controladas (experimentales) o no controladas. Así, los modelos de sobrevivencia se construyen sobre el supuesto de que los tiempos de falla son independientes e idénticamente distribuidos. Sin embargo, en la práctica se observa que las unidades bajo estudio difieren sustancialmente. Para considerar esta heterogeneidad no observada de la población se proponen los modelos univariados de vulnerabilidad (en inglés, *frailty models*) en el análisis de sobrevivencia.

La idea central es que las unidades poseen diferentes grados de vulnerabilidad, y que aquellas que son más vulnerables o frágiles alcanzarán el evento de interés más pronto que las otras. En consecuencia, se podría hacer una selección de las unidades más robustas al evento (con vulnerabilidad baja). Para abordar el problema de la heterogeneidad no observada en los tiempos de eventos resultantes de covariables no observadas, Vaupel, Manton y Stallard [69] y Lancaster [42], de forma independiente, sugirieron un modelo de efectos aleatorios para el tiempo hasta el evento o la duración.

El concepto de fragilidad o vulnerabilidad suministra una forma conveniente de introducir la heterogeneidad y la asociación no observada en los modelos de sobrevida clásicos. En su forma más simple, la fragilidad es un factor aleatorio no observable que modifica la función de riesgo de una unidad o de las unidades relacionadas.

Los modelos de vulnerabilidad univariados extienden el modelo de Cox (6.2.1) de tal forma que el riesgo de una unidad depende, además, de una variable aleatoria no observable Z, la cual actúa multiplicativamente sobre la función de riesgo base, $h_0(t)$. El modelo de vulnerabilidad o de fragilidad para la i-ésima unidad es:

$$h(t|Z_i, X_i) = h_0(t)\exp\{\phi W_i + X_i'\beta\} = Z_i h_0(t)\exp\{X_i'\beta\}. \qquad (7.5.1)$$

La fragilidad $Z_i = \exp\{\phi W_i\}$ es una variable aleatoria, en general no negativa, por lo que el riesgo individual se disminuye cuando $Z_i < 1$ y aumenta en el caso contrario. De otra manera, cuando los Z toman valores mayores que 1, las respectivas unidades tienden a fallar con una tasa más rápida que la tasa de falla con que lo harían estas unidades bajo un modelo

de independencia con los Z_i iguales a 1. Recíprocamente, cuando los Z_i son menores que 1, las unidades tienden a tener una sobrevivencia mayor que la que tendrían bajo un modelo de independencia.

En un modelo de fragilidad (7.5.1), se supone que la función de riesgo puede separarse en componentes multiplicativos: la fragilidad, Z, la función de riesgo de referencia o base, $h_0(t)$, y el predictor lineal, $\exp\{X'\beta\}$.

Así, el modelo de vulnerabilidad es un modelo para la variación individual. Se observa que cuando $\phi = 0$ el modelo (7.5.1) se reduce al modelo de Cox (6.2.1). La respectiva función de sobrevivencia, $S(\cdot)$, describe la fracción de unidades en riesgo al evento en la población objeto del estudio. Esta viene dada por

$$S(t|Z, X) = \exp\left\{-Z\exp[X'\beta]\int_0^t h_0(s)ds\right\} = \{S_0(t)\}^{-Z\exp[X'\beta]}. \tag{7.5.2}$$

La función de sobrevivencia, $S(t|Z, X)$, se puede interpretar como la fracción de unidades que al tiempo t, después de iniciar el seguimiento, no han alcanzado el evento, dado el vector de covariables observables X y la vulnerabilidad Z.

El modelo de riesgos proporcionales de Cox se obtiene de (7.5.1) si la distribución de fragilidad degenera a $Z = 1$ para todas las unidades.

El término *fragilidad* puesto en este modelo, además de considerar e incluir la heterogeneidad de las unidades, permite también evaluar el efecto de las covariables que por alguna razón no fueron observadas y agregadas al modelo.

La función de verosimilitud de los datos $(T_i, \delta_i, X_i, Z_i)$, para $i = 1, \ldots, n$, es similar a las expresiones (6.3.1) y (6.3.2) y está dada por

$$\prod_{i=1}^{n}\left[Z_i h_0(t_i)\exp\{X_i'\beta\}\right]^{\delta_i}(S_0(t_i))^{Z_i\exp\{X_i'\beta\}}. \tag{7.5.3}$$

Entre las distribuciones de probabilidad que se pueden considerar para la variable aleatoria de la fragilidad, Z, la más empleada es la distribución gama con media igual a 1. Así, si la variable aleatoria $Z \sim \Gamma(\eta, \nu)$, con $\eta, \nu \geq 0$, tomando $\eta = \nu = \xi^{-1}$, entonces $E(Z) = 1$ y $var(Z) = \xi$.

La varianza de la variable de fragilidad, ξ, puede ser considerada como una manera de cuantificar la variabilidad presente en las unidades. De manera que si $\xi = 0$, toda la variable de fragilidad será igual a 1, es decir que la distribución gama se ha degenerado al punto 1, con lo que se obtiene un modelo de riesgos proporcionales de Cox para datos independientes.

7.5.0.1. Modelo semiparamétrico de fragilidad gama

En el modelo de fragilidad semiparamétrico no se asume forma alguna de la función de riesgo base. Se requiere de nuevas estrategias de estimación en comparación con las presentadas anteriormente. A continuación, se presenta la extensión del modelo semiparamétrico de riesgos proporcionales (modelo de Cox) al modelo de fragilidad gama semiparamétrico. La función de verosimilitud es similar a la correspondiente al modelo paramétrico, pero ahora la función de riesgo base, $h_0(t)$, se trata como un ruido. En lo que sigue se explica el algoritmo EM (Esperanza Máxima), pues este permite la estimación de los parámetros en el *modelo de fragilidad semiparamétrico gama*. Este algoritmo fue propuesto por Dempster, Lair y Rudin [19] y a menudo se usa en presencia de datos faltantes. El algoritmo EM fue adoptado primero por Nielsen *et al.* [58] para la estimación de parámetros en modelos de fragilidad.

El algoritmo EM itera entre dos pasos. En el primero, se estiman las esperanzas de las fragilidades no observadas basadas en los datos observados, y se obtienen estimaciones usuales. Estas estimaciones ahora se usan en el paso de maximización para obtener nuevas estimaciones de parámetros, dadas las fragilidades estimadas.

Un modelo semiparamétrico de fragilidad se expresa en la forma

$$h(t|X_i, \mathbf{Z}) = \mathbf{Z}h_0(t)\exp\{X_i'\beta\}, \tag{7.5.4}$$

las fragilidades $\mathbf{Z}$ se consideran como variables aleatorias con distribución gama de media igual a 1 y varianza desconocida ξ, es decir, $Z \sim \Gamma(1/\xi, 1/\xi)$. La varianza ξ se puede considerar en este modelo como una medida de la heterogeneidad presente en las unidades.

La estimación se basa en la construcción de una función de verosimilitud y su respectiva optimización respecto a los parámetros que la identifican. El algoritmo EM es empleado como método de optimización, considerando que los valores de la fragilidad faltan o están perdidos.

7.5.0.2. Estimación vía algoritmo EM

Primero, se considera la verosimilitud completa, con las variables de fragilidad asumidas como variables aleatorias observadas, similares a los tiempos al evento. Esta verosimilitud completa resulta de la densidad conjunta de

(t_i, δ_i, Z_i) para $i = 1, \ldots, n$. Esta es de la forma

$$\begin{aligned} L(\beta, \xi|\mathbf{Z}) &= \prod_{i=1}^{n} f(t_i, \delta_i, Z_i, \beta, \xi) \\ &= \prod_{i=1}^{n} f(t_i, \delta_i, Z_i, \beta) \prod_{i=1}^{n} f(Z_i, \xi) \\ &= L_1(\beta|\mathbf{Z}) L_2(\xi|\mathbf{Z}), \end{aligned} \tag{7.5.5}$$

con $\mathbf{Z} = (Z_1, \ldots, Z_n)$ y

$$L_1(\beta|\mathbf{Z}) = \prod_{i=1}^{n} \left(Z_i h_0(t_i) e^{X_i'\beta}\right)^{\delta_i} (S_0(t_i))^{Z_i e^{X_i'\beta}}. \tag{7.5.6}$$

El segundo término viene dado por la densidad de probabilidad de las variables de fragilidad, este es:

$$L_2(\xi|\mathbf{Z}) = \prod_{i=1}^{n} f(Z_i, \xi). \tag{7.5.7}$$

Si las fragilidades Z_i fueran conocidas, los parámetros de regresión β podrían ser estimados mediante el método de verosimilitud parcial reescribiendo los términos $Z_i e^{X_i\beta}$ en (7.5.6), de la forma $e^{X_i\beta + \ln(Z_i)}$, usando $\ln(Z_i)$ como un valor *offset* fijo. Por consiguiente, el paso de la esperanza es necesario para obtener valores de estimación de las fragilidades. Estas estimaciones se usan en lugar de las fragilidades desconocidas en el paso de maximización para obtener las estimaciones de los parámetros de regresión.

Paso de maximización:

Se extiende la idea de verosimilitud parcial en el modelo de Cox contenida en la expresión (6.3.6). La verosimilitud parcial para los parámetros de regresión en el modelo de fragilidad es:

$$L(\beta|\mathbf{Z}) = \prod_{i=1}^{n} \left(\frac{\exp\left\{\mathbf{x}_i'\beta + \ln(Z_i)\right\}}{\sum_{l \in \mathcal{R}_i} Z_{(l)} \exp\left\{\mathbf{x}_{(l)}'\beta\right\}} \right)^{\delta_i}. \tag{7.5.8}$$

Las variables aleatorias Z_i y $\ln(Z_i)$ se sustituyen ahora por sus valores esperados (en la k-ésima iteración) $E_{(k)}(Z_i)$ y $E_{(k)}(\ln Z_i)$. Tomando logaritmos, se tiene

$$L(\beta, \xi) = \sum_{i=1}^{n} \delta_i \left[\mathbf{x}_i'\beta + E_{(k)}(\ln(Z_i)) - \ln\left(\sum_{l \in \mathcal{R}_i} E_{(k)}(Z_{(l)}) \exp\left\{\mathbf{x}_{(l)}'\beta\right\} \right) \right]. \tag{7.5.9}$$

A partir de esta expresión, se pueden obtener nuevas estimaciones (la k-ésima), $\beta_{(k)}$, usando una herramienta computacional adecuada como R o SAS. Una nueva estimación del parámetro de fragilidad, $\xi_{(k)}$, es obtenida por maximización de $L_2(\xi|\mathbf{Z})$ y también reemplazando las variables desconocidas Z_i por sus valores esperados en la iteración o k-ésima etapa.

Paso Valor Esperado (esperanza):

Los valores no observados de fragilidad Z_i para cada unidad, $i = 1, \ldots, n$, pueden estimarse mediante la expresión

$$E_{(k+1)}(Z_i) = \frac{1/\xi_{(k)} + \delta_i}{1/\xi_{(k)} + \tilde{H}_{(k)}(t_i)e^{\mathbf{x}_i'\beta_{(k)}}}. \tag{7.5.10}$$

En esta última expresión $\tilde{H}_{(k)}(\cdot)$ es un estimador de la función de riesgo acumulado base, $H_0(t)$, evaluada sobre los parámetros estimados en la k-ésima iteración. Por ejemplo, el estimador de Nelson-Aalen:

$$\tilde{H}_{NA(k)}(t) = \sum_{i:t_i<t} \frac{\delta_i}{\sum_{l\in\mathcal{R}_i} E_{(k)}(Z_{(l)})e^{\mathbf{x}_{(l)}'\beta_{(k)}}}.$$

Como la distribución Z es gama, $\ln(Z)$ tiene una distribución log-gama con valor esperado

$$E_{(k+1)}(\ln(Z_i)) = \psi(1/\xi + \delta_i) - \ln(1/\xi + \tilde{H}_{(k)}(t_i)e^{\mathbf{x}_i'\beta_{(k)}}),$$

con la función digama $\psi(x) = \Gamma'(x)/\Gamma(x)$.

Ejemplo 7.5.1. Hanagal [30, p. 6] muestra los resultados de un ensayo clínico de un medicamento, 6-mercaptopurina (6-MP), frente a un placebo, suministrados a 21 niños que padecen leucemia aguda.

Los pacientes fueron seleccionados entre quienes tuvieron una remisión completa o parcial de su leucemia inducida por el tratamiento con el medicamento prednisona. Una remisión completa o parcial significa que todos o la mayoría los signos de enfermedad han desaparecido de la médula ósea. El ensayo fue conducido por pares de pacientes de un hospital, clasificados por estado de remisión (completo o parcial) y asignados aleatoriamente dentro de cada par, a uno de los cuales se le suministró 6-MP y al otro un placebo. Los pacientes fueron seguidos hasta que su leucemia reapareció (recaída) o hasta el final del estudio (en meses).

Tabla 7.7. Duración de la remisión en pacientes con leucemia aguda

Par	S	TP	T6	RI	Pair	S	TP	T6	RI
1	1	1	10	1	12	1	5	20	0
2	2	22	7	1	13	2	4	19	0
3	2	3	32	0	14	2	15	6	1
4	2	12	23	1	15	2	8	17	0
5	2	8	22	1	16	1	23	35	0
6	1	17	6	1	17	1	5	6	1
7	2	2	16	1	18	2	11	13	1
8	2	11	34	0	19	2	4	9	0
9	2	8	32	0	20	2	1	6	0
10	2	12	25	0	21	2	8	10	0
11	2	2	11	0					

Las variables contenidas en cada una de las columnas de la tabla 7.7 son las siguientes:

- Par.
- Estado de la remisión en la aleatorización (S) (1 = parcial, 2 = completa).
- Tiempo de recaída para pacientes con placebo, meses (TP).
- Tiempo de recaída para pacientes de 6-MP, meses (T6).
- Indicador de recaída (RI) (0 = censurado, 1 = recaída) para pacientes de 6-MP.

El código R, dentro del paquete *survival*, junto con los respectivos resultados del procesamiento, se transcriben a continuación.

```
 #MODELO DE COX ORDINARIO
cox<-coxph(Surv(TPO, RI) ~ S, data=leucemia)
summary(cox)
Call:
coxph(formula = Surv(TPO, RI) ~ S, data = leucemia)
n= 42, number of events= 30
```

```
coef exp(coef) se(coef)    z Pr(|z|)
S -0.2070    0.8130   0.4143 -0.5    0.617
exp(coef) exp(-coef) lower .95 upper .95
S     0.813       1.23     0.361     1.831
Concordance= 0.533  (se = 0.045 )
Rsquare= 0.006   (max possible= 0.988 )
Likelihood ratio test= 0.24  on 1 df,   p=0.6231
Wald test            = 0.25  on 1 df,   p=0.6174
Score (logrank) test = 0.25  on 1 df,   p=0.6168

cox$loglik
[1] -93.18427 -93.06351
#MODELO DE FRAGILIDAD DE COX
coxfrag<-coxph(Surv(TPO, RI) ~ S+frailty(Par), data=leucemia)
summary(coxfrag)
Call:
coxph(formula = Surv(TPO, RI) ~ S + frailty(Par), data = leucemia)
n= 42, number of events= 30
coef   se(coef) se2    Chisq DF p
S            -0.207 0.4143   0.4143 0.25  1  0.62
frailty(Par)                        0.00  0  0.95

exp(coef) exp(-coef) lower .95 upper .95
S     0.813       1.23     0.361     1.831

Iterations: 6 outer, 23 Newton-Raphson
Variance of random effect= 5e-07   I-likelihood = -93.1
Degrees of freedom for terms= 1 0
Concordance= 0.615  (se = 0.062 )
Likelihood ratio test= 0.24  on 1 df,   p=0.6231
coxfrag$loglik
[1] -93.18427 -93.06350
```

El log-parcial de la verosimilitud cuando el vector $\beta = 0$ es -93.18427. En el modelo de riesgos proporcionales de Cox, el ajuste final es -93.06351 y en el modelo de Cox de fragilidad gama, la varianza del efecto aleatorio estimada es casi cercana a cero (*Variance of random effect* = 5e-07). Una prueba de razón de verosimilitud para la fragilidad es el doble de la diferencia entre el log de la verosimilitud parcial con el término de fragilidad integrado

fuera, que se muestra como "I-likelihood" en la impresión, y el log de la verosimilitud del modelo de no fragilidad, este es, $2(93.1 - 93.063) = 0.074$.

En el libro de Wienke [72] se extienden los modelos de fragilidad univariados al caso multivariado en los modelos marginales de fragilidad, los modelos de fragilidad compartida y los modelos para datos correlacionados (datos en conglomerados o *cluster*). Finalmente, se incorporan las cópulas para el modelamiento de la fragilidad en datos correlacionados o en conglomerados.

Tabla 7.8. Datos con variable tiempo-dependiente

Pac	ti	tf	Lbrt	cen	Trto	Edad	Lbr
1	0	47	3.2	0	0	46	3.2
1	47	184	3.8	0	0	46	3.2
1	184	251	4.9	0	0	46	3.2
1	251	281	5.0	1	0	46	3.2
2	0	94	3.1	0	0	57	3.1
2	94	187	2.9	0	0	57	3.1
2	187	321	3.1	0	0	57	3.1
2	321	604	3.2	0	0	57	3.1
3	0	61	2.2	0	0	56	2.2
3	61	97	2.8	0	0	56	2.2
3	97	142	2.9	0	0	56	2.2
3	142	359	3.2	0	0	56	2.2
3	359	440	3.4	0	0	56	2.2
3	440	457	3.8	1	0	56	2.2
4	0	92	3.9	0	0	65	3.9
4	92	194	4.7	0	0	65	3.9
4	194	372	4.9	0	0	65	3.9
4	372	384	5.4	1	0	65	3.9
5	0	87	2.8	0	0	73	2.8
5	87	192	2.6	0	0	73	2.8
5	192	341	2.9	0	0	73	2.8
5	341	341	3.4	0	0	73	2.8
6	0	94	2.4	0	0	64	2.4
6	94	197	2.3	0	0	64	2.4
6	197	384	2.8	0	0	64	2.4
6	384	795	3.5	0	0	64	2.4
6	795	842	3.9	1	0	64	2.4
7	0	74	2.4	0	1	69	2.4
7	74	202	2.9	0	1	69	2.4
7	202	346	3.0	0	1	69	2.4
7	346	917	3.0	0	1	69	2.4
7	917	1411	3.9	0	1	69	2.4
7	1411	1514	5.1	1	1	69	2.4
8	0	90	2.4	0	1	62	2.4
8	90	182	2.5	0	1	62	2.4
8	182	182	2.9	0	1	62	2.4
9	0	101	2.5	0	1	71	2.5
9	101	410	2.5	0	1	71	2.5
9	410	774	2.7	0	1	71	2.5
9	774	1043	2.8	0	1	71	2.5
9	1043	1121	3.4	1	1	71	2.5
10	0	182	2.3	0	1	69	2.3
10	182	847	2.2	0	1	69	2.3
10	847	1051	2.8	0	1	69	2.3
10	1051	1347	3.3	0	1	69	2.3
10	1347	1411	4.9	0	1	69	2.3
11	0	167	3.8	0	1	77	3.8
11	167	498	3.9	0	1	77	3.8
11	498	814	4.3	1	1	77	3.8
12	0	108	3.1	0	1	58	3.1
12	108	187	2.8	0	1	58	3.1
12	187	362	3.4	0	1	58	3.1
12	362	694	3.9	0	1	58	3.1
12	694	1071	3.8	1	1	58	3.1

Tabla 7.9. Tiempos de sobrevida de pacientes con cáncer de mama

izq	der	terap	cens	izq	der	terap	cens
0	7	1	1	0	22	0	1
0	8	1	1	0	5	0	1
0	5	1	1	4	9	0	1
4	11	1	1	4	8	0	1
5	12	1	1	5	8	0	1
5	11	1	1	8	12	0	1
6	10	1	1	8	21	0	1
7	16	1	1	10	35	0	1
7	14	1	1	10	17	0	1
11	15	1	1	11	13	0	1
11	18	1	1	11	NA	0	0
15	NA	1	0	11	17	0	1
17	NA	1	0	11	NA	0	0
17	25	1	1	11	20	0	1
17	25	1	1	12	20	0	1
18	NA	1	0	13	NA	0	0
19	35	1	1	13	39	0	1
18	26	1	1	13	NA	0	0
22	NA	1	0	13	NA	0	0
24	NA	1	0	14	17	0	1
24	NA	1	0	14	19	0	1
25	37	1	1	15	22	0	1
26	40	1	1	16	24	0	1
27	34	1	1	16	20	0	1
32	NA	1	0	16	24	0	1
33	NA	1	0	16	60	0	1
34	NA	1	0	17	27	0	1
36	44	1	1	17	23	0	1
36	48	1	1	17	26	0	1
36	NA	1	0	18	25	0	1
36	NA	1	0	18	24	0	1
37	44	1	1	19	32	0	1
37	NA	1	0	21	NA	0	0
37	NA	1	0	22	32	0	1
37	NA	1	0	23	NA	0	0
38	NA	1	0	24	31	0	1
40	NA	1	0	24	30	0	1
45	NA	1	0	30	34	0	1
46	NA	1	0	30	36	0	1
46	NA	1	0	31	NA	0	0
46	NA	1	0	32	NA	0	0
46	NA	1	0	32	40	0	1
46	NA	1	0	34	NA	0	0
46	NA	1	0	34	NA	0	0
46	NA	1	0	35	NA	0	0
46	NA	1	0	35	39	0	1
				44	48	0	1
				48	NA	0	0

Capítulo

ocho

Modelos de riesgo aditivos

$$h(t|X(t)) = \mathbb{X}(t)'\beta(t)$$

8.1. Introducción

En el modelo de riesgos proporcionales, el efecto de las covariables es actuar multiplicativamente sobre una tasa de riesgo base. Las covariables que no tengan efectos multiplicativos sobre la tasa de riesgo base se modelan, bien sea por la inclusión de una covariable dependiente del tiempo, o bien sea modelando por estratos (niveles de las covariables). Una alternativa es un modelo de riesgo aditivo. Como en el modelo multiplicativo, se tiene una variable aleatoria, $\mathbf{T}$, de *tiempo hasta el evento*, cuya distribución depende de un vector de, posiblemente, covariables tiempo-dependientes, $\mathbf{X}(t) = (X_1(t), \ldots, X_p(t))$. Se asume que la tasa de riesgo al tiempo t, para la unidad i, con $i = 1, \ldots, n$, identificada por el vector de covariables $\mathbf{X}_i(t)$, es una combinación lineal de las $X_j(t)$. Esta es

$$h_i(t|\mathbf{X}(t)) = \beta_0(t) + \sum_{j=1}^{p} \beta_j(t) X_{ij}(t), \text{ para } i = 1, \ldots, n,$$

donde $\beta_j(t)$ son los parámetros funcionales dependientes del tiempo; es decir, el efecto de las covariables $X_j(t)$ es también dependiente del tiempo, las cuales deben estimarse desde los datos.

Dentro de la estructura del modelo 8.1 se pueden considerar dos tipos de modelos. El primero, propuesto por Aalen [1], permite que los coeficientes de regresión, $\beta_j(t)$, sean funciones cuyos valores pueden cambiar en el tiempo. Este modelo permite que las covariables fijas en el tiempo tengan un efecto dinámico o cambiante en el tiempo sobre el riesgo al evento. El segundo modelo, debido a Lin y Ying [47], reemplaza las funciones de regresión $\beta_j(t)$ por constantes β_j. El modelo toma la forma

$$h_i(t|\mathbf{X}(t)) = \beta_0(t) + \sum_{j=1}^{p} \beta_j X_{ij}(t), \text{ para } i = 1, \ldots, n.$$

8.2. Modelo de riesgo aditivo de Aalen no paramétrico

En los estudios longitudinales las unidades son observadas a lo largo del tiempo para verificar la ocurrencia de un determinado evento. Usualmente se puede asumir que la ocurrencia del evento es independiente entre las unidades. Además se asume que la distribución del tiempo, $\mathbf{T}_i$, para la ocurrencia del evento sobre la unidad $i = 1, \ldots, n$, depende del vector de covariables que identifican la unidad $i = 1, \ldots, n$, $\mathbf{X}_i(t) = (X_{i1}(t), \ldots, X_{ip}(t))$,

posiblemente dependientes del tiempo. El modelo de riesgos aditivos de Aalen está dado por:

$$h_i(t|\mathbf{X}(t)) = \beta_0(t) + \sum_{j=1}^{p} \beta_j(t) X_{ij}(t), \text{ para } i = 1, \ldots, n. \tag{8.2.1}$$

Este modelo escrito en forma matricial toma la forma

$$h(t|\mathbf{X}(t)) = \mathbb{X}(t)'\beta(t), \tag{8.2.2}$$

donde $\beta(t) = (\beta_0(t), \beta_1(t), \ldots, \beta_p(t))'$ es un vector de funciones del tiempo desconocidas, las cuales deben ser estimadas. El funcional $\beta_0(t)$ se puede interpretar como una función de riesgo base (similar a $h_0(t)$ en el modelo de Cox). Las funcionales $\beta_j(t)$, $j = 1, \ldots, p$, llamadas funciones de regresión, miden el efecto de las respectivas covariables de manera dinámica en el tiempo. La matriz, $\mathbb{X}(t)$, de tamaño $n \times (p + 1)$, se define de la siguiente manera: si el evento considerado aún no ocurre para la i-ésima unidad y no está censurada, entonces la i-ésima fila de $\mathbb{X}(t)$ es un vector de la forma $\mathbf{X}_i(t) = (1, X_{i1}(t), X_{i2}(t), \ldots, X_{ip}(t))'$. En el caso contrario, es decir, si la unidad no está en riesgo al tiempo t, entonces la fila correspondiente de $\mathbb{X}(t)$ es una fila de ceros. A manera de ilustración, considere que se tienen $n = 5$ unidades y $p = 3$ covariables. Al inicio del estudio, es decir, con $t = 0$, todas las unidades están en riesgo al evento, por lo que la matriz de covariables es

$$\mathbb{X}(0) = \begin{pmatrix} 1 & X_{11} & X_{12} & X_{13} \\ 1 & X_{21} & X_{22} & X_{23} \\ 1 & X_{31} & X_{32} & X_{33} \\ 1 & X_{41} & X_{42} & X_{43} \\ 1 & X_{51} & X_{52} & X_{53} \end{pmatrix}.$$

Si en el tiempo $t = t_1 > 0$ solamente estuviesen en riesgo al evento las unidades 1, 4 y 5, entonces la matriz con los valores de las covariables en este tiempo estaría dada por

$$\mathbb{X}(t_1) = \begin{pmatrix} 1 & X_{11} & X_{12} & X_{13} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & X_{41} & X_{42} & X_{43} \\ 1 & X_{51} & X_{52} & X_{53} \end{pmatrix}.$$

Este modelo se considera no paramétrico, en el sentido de que ninguna forma paramétrica particular se asume para las funciones de regresión $\beta_j(t)$.

Estas funciones pueden variar arbitrariamente en el tiempo, mostrando cambios en el efecto de la covariables sobre el riesgo al evento.

En contraste con el modelo de riesgos proporcionales de Cox, donde la técnica de estimación se hace a través de la verosimilitud, la estimación de las funciones de regresión o coeficientes de riesgo se basan en las técnicas de mínimos cuadrados.

Los datos consisten en una muestra $(\mathbf{T}_i, \delta_i, \mathbb{X}_i(t))$, $i = 1, \ldots, n$, donde $\mathbf{T}_i$ es el tiempo para el evento sobre la i-ésima unidad, δ_i es el indicador del evento, y $\mathbb{X}_i(t) = (X_{i1}(t), \ldots, X_{ip}(t))$ es un vector de covariables posiblemente dependientes del tiempo.

8.3. Estimación

En el modelo de riesgos proporcionales de Cox, los coeficientes de regresión β, que representan los efectos multiplicativos sobre la función de riesgo, son constantes desconocidas cuyos valores no cambian en el tiempo. Para el modelo aditivo de Aalen, se asume que las covariables actúan de manera aditiva sobre la función de riesgo base. El efecto se modela a través de las funciones de regresión desconocidas $\beta_j(t)$, denominadas también como *coeficientes de riesgo*, que son funciones del tiempo, es decir, el efecto de las covariables puede cambiar durante el desarrollo del estudio.

La estimación directa de los $\beta_j(t)$'s es difícil en la práctica. La derivación de estos estimadores se basa en la aproximación al análisis de sobrevida mediante un proceso de conteo (7.2.0.1), y es similar a la derivación del estimador para la función de riesgo acumulado de Nelson-Aalen (sección 2.5).

Considérese el vector columna $\mathbf{B}(t)$, cuyos elementos $B_j(t)$, $j = 1, \ldots, p$ corresponden a las funciones de riesgo acumulado, definidas por

$$B_j(t) = \int_0^t \beta_j(u)du, \;\; j = 1, \ldots, p. \tag{8.3.1}$$

Las estimaciones empíricas de $\beta_j(t)$ son dadas por estimaciones de la pendiente del estimador de $B_j(t)$. Esta afirmación se hace a partir del concepto de la derivada de una función contenido en la expresión (8.3.1).

Para encontrar el estimador de $B_j(t)$ se emplea la técnica de mínimos cuadrados. Para este propósito se redefine y estructuran los vectores y matrices contenidas en el modelos (8.2.2). Sea

$$Y_i(t) = \begin{cases} 1, & \text{si la unidad } i \text{ esta bajo observación (en riesgo) al tiempo } t \\ 0, & \text{si la unidad } i \text{ no esta bajo observación (no riesgo) al tiempo } t. \end{cases}$$

Para datos con censura a derecha $Y_i(t)$, es 1 si $t \leq \mathbf{T}_i$.

Se redefine la matriz de diseño, $\mathbb{X}_a(t)$, de tamaño $n \times (p+1)$, cuya i-ésima fila es $\mathbb{X}_{ai}(t) = Y_i(t)\mathbb{X}_i(t))$. Es decir, $\mathbb{X}_{ai} = (1, X_{i1}(t), \ldots, X_{ip}(t))$ si la i-ésima unidad está en riesgo al tiempo t, y un vector de $(p+1)$ ceros si la unidad no está en riesgo. El caso anterior de $n = 5$ unidades y $p = 3$ covariables es un ejemplo de estas matrices; así, $\mathbb{X}(t_1) = \mathbb{X}_a(t)$.

Sea $\mathbf{I}(t)$ un vector de tamaño $n \times 1$ con el k-ésimo elemento igual a 1, si a la k-ésima unidad le ha ocurrido el evento, y 0 en el otro caso. El estimador de mínimos cuadrados del vector $\mathbf{B} = (B_0(t), B_1(t), \ldots, B_p(t))'$ es

$$\widehat{\mathbf{B}}(t) = \sum_{T_i \leq t} [\mathbb{X}'_{ai}(T_i)\mathbb{X}_{ai}(T_i)]^{-1}\mathbb{X}'_{ai}(T_i)\mathbf{I}(T_i). \tag{8.3.2}$$

La matriz de varianza-covarianzas estimada de

$$\widehat{Var}(\widehat{\mathbf{B}}(t)) = \sum_{T_i \leq t} [\mathbb{X}'_{ai}(T_i)\mathbb{X}_{ai}(T_i)]^{-1}\mathbb{X}'_{ai}(T_i)\mathbf{I}^D(T_i)\mathbb{X}_{ai}(T_i)\{[\mathbb{X}'_{ai}(T_i)\mathbb{X}_{ai}(T_i)]^{-1}\}'. \tag{8.3.3}$$

La matriz $\mathbf{I}^D(t)$ es la matriz diagonal con elementos iguales a $\mathbf{I}(t)$. El estimador $\widehat{\mathbf{B}}(t)$ solo existe hasta el tiempo t; este es el tiempo más pequeño en el cual $[\mathbb{X}'_{ai}(T_i)\mathbb{X}_{ai}(T_i)]$ es singular.

El estimador $\widehat{B}_j$ estima la integral de regresión de (8.3.1) en la misma forma que el estimador de Nelson-Aalen (sección 2.5), es decir, estima la integral de la tasa de riesgo en el caso univariado. De igual manera a como se construyen bandas de confianza para las funciones de riesgo acumulado, se construyen bandas de confianza para B_j.

Con los resultados anteriores, y con base en los valores de las covariables, se puede obtener una estimación del riesgo acumulado al tiempo t. Así, sea $X_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip)}$ el conjunto de las covariables fijadas en el tiempo cero para la i-ésima unidad. Un estimador del riesgo acumulado $\widehat{H}(t)$ está dado por:

$$\widehat{H}_i(t) = X'\widehat{\mathbf{B}}(t) = \widehat{B}_0(t) + \sum_{\substack{j=1 \\ t \leq \tau}}^{p} x_{ij}\widehat{B}_j(t), \tag{8.3.4}$$

con $\widehat{B}_j(t)$ siendo el estimador de mínimos cuadrados dado por (8.3.2). Estas estimaciones se tienen para $t \leq \tau$, con τ como el valor maximal de t para el cual una matriz es no singular.

Una aproximación a un intervalo de $(1-\alpha)100\,\%$ de confianza para B_j está dado por

$$\widehat{B}_j(t) \mp Z_{1-\alpha/2}\sqrt{\widehat{Var}[\widehat{B}_j(t)]},\ j = 1, \ldots, p+1.$$

8.3.0.1. Efecto de las covariables

Una vez que se han estimados los componentes de un modelo, el interés se dirige a verificar si una covariable específica tiene algún efecto en la función de riesgo. Para el modelo aditivo de Aalen, esto corresponde a verificar la hipótesis nula

$$H_{0j}:\ \beta_j(t) = 0, \qquad t \in [0, \tau],\ j = 1, \ldots, p+1.$$

Para desarrollar la verificación o contraste de la hipótesis, se necesita una matriz de ponderaciones en la construcción de la estadística. La matriz de ponderación es una matriz diagonal, $\mathbf{W}(t)$, con elementos en la diagonal, $\mathbf{W}_j(t)$, $j = 1, \ldots, p+1$. Con estas ponderaciones se construye la estadística

$$\mathbf{U} = \sum_{T_i} \mathbf{W}(T_i)[\mathbb{X}'_{ai}(T_i)\mathbb{X}_{ai}(T_i)]^{-1}\mathbb{X}'_{ai}(T_i)\mathbf{I}(T_i). \tag{8.3.5}$$

El elemento $(j+1)$-ésimo de $\mathbf{U}$ es la estadística de verificación empleada para verificar $H_0:\ \beta_j(t) = 0$.

La matriz de covarianzas de $\mathbf{U}$ es dada por

$$\mathbf{V} = \sum_{T_i} \mathbf{W}(T_i)[\mathbb{X}'_{ai}(T_i)\mathbb{X}_{ai}(T_i)]^{-1}\mathbb{X}'_{ai}(T_i)\mathbf{I}^D(T_i)\mathbb{X}_{ai}(T_i)[\mathbb{X}'_{ai}(T_i)\mathbb{X}_{ai}(T_i)]^{-1}\mathbf{W}(T_i). \tag{8.3.6}$$

Los elementos $\mathbf{U}$ son sumas ponderadas de los incrementos de $\widehat{B}_j(t)$ y los elementos de $\mathbf{V}$ se obtienen también desde los elementos de $\widehat{Var}(\widehat{\mathbf{B}}(t))$.

Usando estas estadísticas conjuntamente, se verifica la hipótesis de que $\beta_j(t) = 0$ para todo $j \in \mathbf{J}$, el cual es un subconjunto de $\{0, 1, \ldots, p+1\}$, mediante la estadística

$$\mathbb{V} = \mathbf{U}'_j\mathbf{V}_j^{-1}\mathbf{U}_j, \tag{8.3.7}$$

donde $\mathbf{U}_j$ es el subvector de $\mathbf{U}$ correspondiente a los elementos en $\mathbf{J}$, es decir, el número de coeficientes de riesgo, y $\mathbf{V}_j$ es la correspondiente submatriz de covarianzas en $\mathbf{V}$. Aalen sugiere el ponderador $\mathbf{W}(t) = \{diag([\mathbb{X}(t)'\mathbb{X}(t)])\}^{-1}$.

La estadística $\mathbb{V}$ tiene una distribución asintótica ji-cuadrado con s grados de libertad, donde s es el número de componentes del subconjunto de $\beta_j(t)$; es decir, $1 \leq s \leq (p+1)$.

Para diagnosticar el modelo aditivo de Aalen tal como en los modelos anteriores, se pueden emplear los residuales con algunas modificaciones dadas por la estructura del modelo de Aalen.

Ejemplo 8.3.1. Se retoma el caso de cáncer de laringe descrito y desarrollado en el ejemplo 6.4.2.5 con el modelo de Cox. Las covariables registradas en el diagnóstico para cada paciente fueron el estadio de la enfermedad (I, II, III o IV) y la edad (en años), la cual fue centrada en su media.

El modelo propuesto es el modelo de regresión de la función de riesgo $h(t)$ en función de las covariables *edad* y *estadio de la enfermedad* (*est*). Se escribe como

$$h(t|edad\_c,\ est) = \beta_0(t) + \beta_1(t)edad\_c + \beta_2(t)est.$$

Las estimaciones obtenidas a partir del ajuste anterior del modelo aditivo de Aalen se calculan mediante el siguiente código R:

```
#DATOS TABLA 6.7 (se debe estructurar)
laringe<-read.table("Laringe.txt",header=T)
attach(laringe)
#EDAD CENTRADA EN LA MEDIA
edad_c<-laringe$edad-mean(laringe$edad)
laringe1<-cbind(laringe,edad_c)
attach(laringe1)
require(survival)
aalen1<-aareg(Surv(tpo, cens) ~ factor(est)+edad,
x=T,data=laringe1,nmin=1)
aalen1
```

El código "aareg", acrónimo de Aalen Regression, es una función del paquete *survival*. Note que tiene una posición y argumentos similares a "coxph" en la declaración del modelo de función de riesgo proporcional.

La salida del procesamiento con el paquete R se muestra a continuación:

```
> aalen1
Call:
aareg(formula = Surv(tpo, cens)~factor(est)+edad_c,laringe1)

n= 90
```

```
34 out of 34 unique event times used

               slope     coef se(coef)     z        p
Intercept    0.11800 0.010100 0.002870 3.520 0.000426
factor(est)2 0.04670 0.004250 0.005500 0.772 0.440000
factor(est)3 0.18600 0.010300 0.005730 1.800 0.071900
factor(est)4 1.76000 0.039600 0.013600 2.910 0.003620
edad_c       0.00333 0.000283 0.000262 1.080 0.280000

Chisq=11.63 on 4 df, p=0.02; test weights=aalen
```

En estos resultados, se observa que la covariable *edad*, centrada en su media, no es significativa en la función de riesgo (el valor p es igual a 0.280). Así, esta variable se descarta del modelo.

El valor de la estadística (8.3.7) es igual a 11.63, el cual, para una distribución ji-cuadrado con $s = 4$ grados de libertad, tiene un valor p igual a 0.02, con lo cual se puede afirmar que al menos una de las funciones de regresión (coeficientes de riesgo) es significativamente distinta de cero.

El modelo aditivo de Aalen ajustado para estos datos tiene como única covariable el estadio de la enfermedad (*est*). El código R para obtener las estimaciones de este modelo es:

```
aalen2<-aareg(Surv(tpo, cens) ~ factor(est),laringe1)
```

Los resultados del procesamiento del modelo de riesgo

$$h(t|edad\_c,\ est) = \beta_0(t) + \beta_1(t)est \tag{8.3.8}$$

son los siguientes:

```
> aalen2
Call:
aareg(formula = Surv(tpo, cens) ~ factor(est), data = laringe1)

n= 90
34 out of 34 unique event times used

              slope     coef se(coef)    z        p
Intercept    0.1270 0.011800  0.00305 3.87 0.000108
factor(est)2 0.0197 0.000897  0.00529 0.17 0.865000
factor(est)3 0.1790 0.009620  0.00575 1.67 0.094600
factor(est)4 1.7500 0.039100  0.01360 2.88 0.003980
```

```
Chisq=10.78 on 3 df, p=0.013; test weights=aalen
```

La calidad del modelo aditivo de Aalen, evaluado con la única covariable *estadio* (*est*), se determina mediante un gráfico de residuales asociados con el modelo ajustado a los datos.

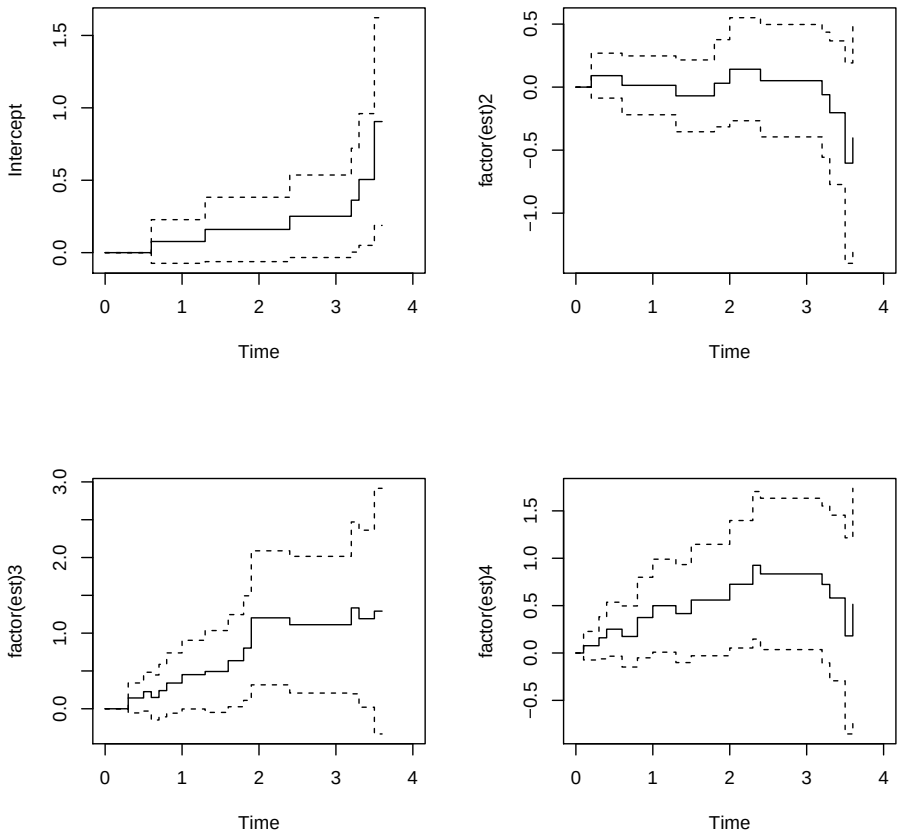

Figura 8.1. Funciones de regresión acumulada del modelo 8.3.8.

Para el modelo ajustado (8.3.8), las funciones de regresión acumulada, junto con sus respectivas bandas de confianza, se muestran en la figura 8.1. Estos gráficos se generan con el código R siguiente:

```
par(mfrow=c(2,2))
plot(aalen2,xlim=c(0,4))
```

El panel de la parte superior izquierda muestra la estimación del riesgo acumulado base, representado por $\widehat{B}_0(t)$, que corresponde a los pacientes en estadio I. A partir de este panel se puede observar que el riesgo crece gradualmente a lo largo del tiempo de observación y se acentúa después del tercer año. Los demás paneles muestran un crecimiento del riesgo acumulado a la muerte de los pacientes en los estadios II, III y IV en relación con el riesgo acumulado de los pacientes en el estadio I. El panel superior derecho presenta una inclinación cercana a cero, lo cual significa que pacientes en el estadio II presentan un riesgo muy similar al de aquellos pacientes observados en estadio I.

Los paneles de la parte inferior muestran inclinaciones acentuadas durante los dos primeros años. Este hecho muestra que los pacientes en estadio III o en el estadio IV presentan, en comparación con los pacientes en estadio I, un riesgo más elevado en los dos primeros años. Después de estos dos primeros años la diferencia de esos riesgos acumulados permanece casi constante (inclinación próxima a cero) y con un crecimiento gradual para los pacientes observados en estadio IV en comparación con aquellos que están en el estadio I de su enfermedad.

8.4. Modelo de riesgo aditivo de Lin-Ying

En el modelo de riesgo aditivo de Aalen, los coeficientes de regresión se consideran como funciones del tiempo. El modelo, como su nombre lo indica, se debe a Lin y Ying [47], [48] y [49], en adelante modelo LY. Ellos proponen una alternativa de modelo de regresión para el riesgo aditivo. En el modelo de LY las funciones de regresión del modelo de Aalen se reemplazan por constantes. Así, el modelo aditivo de LY para la tasa de riesgo condicional de una unidad i identificada por las covariables $\mathbb{X}_i(t)$ es

$$h(t|\mathbb{X}_i(t)) = \alpha_0(t) + \sum_{j=1}^{p} \alpha_j X_{ij}(t), \tag{8.4.1}$$

donde α_j, $j = 1, \ldots, p$ son parámetros desconocidos y $\alpha_0(t)$ es una función arbitraria de base. Cuando todas las covariables son fijadas en el tiempo 0, es fácil estimar los coeficientes de regresión α_j, $j = 1, \ldots, p$.

Los datos consisten en una muestra $(\mathbf{T}_i, \delta_i, \mathbb{X}_i)$, para $i = 1, \ldots, n$, donde $\mathbf{T}_i$ es el tiempo para el evento, δ_i es el indicador del evento, y $\mathbb{X}_i = (X_{i1}, \ldots, X_{ip})$ es un vector de tamaño $p \times 1$ de posibles covariables dependientes del tiempo. Se asume que los $\mathbb{T}_i$ son ordenados de menor a mayor con $0 = T_0 \leq T_i \leq T_2 \leq \cdots\cdots T_n$.

Se define para la i-ésima unidad

$$Y_i(t) = \begin{cases} 1, & \text{si la unidad } i \text{ está en observación (riesgo) al tiempo } t \\ 0, & \text{si la unidad } i \text{ no está en observación (no riesgo) al tiempo } t. \end{cases}$$

Una unidad i con censura a derecha tiene $Y_i(t) = 1$, siempre que $t \leq T_i$.

Para construir los estimadores de α_j, $j = 1, \ldots, p$ se requiere construir el vector $\bar{\mathbf{Z}}(t)$, el cual es el promedio de las covariables al momento t. Este es

$$\bar{\mathbf{Z}}(t) = \frac{\sum_{i=1}^{n} \mathbf{Z}_i(t) Y_i(t)}{\sum_{i=1}^{n} Y_i(t)}. \tag{8.4.2}$$

El numerador es la suma de las covariables para las unidades en riesgo al tiempo t y el denominador es el número de unidades en riesgo al tiempo t.

En la misma línea del proceso de estimación se construye la matriz $\mathbf{A}$, de tamaño $p \times p$, dada por

$$\mathbf{A} = \sum_{k=1}^{n} \sum_{i=1}^{k} (T_i - T_{i-1}) [\mathbf{Z}_k - \bar{\mathbf{Z}}(T_i)]' [\mathbf{Z}_k - \bar{\mathbf{Z}}(T_i)]; \tag{8.4.3}$$

el vector $\mathbf{B}$ $p \times 1$ dado por

$$\mathbf{B}' = \sum_{k=1}^{n} \delta_k [\mathbf{Z}_k - \bar{\mathbf{Z}}(T_k)]; \tag{8.4.4}$$

y la matriz $\mathbf{C}$, de tamaño $p \times p$, definida por

$$\mathbf{C} = \sum_{k=1}^{n} \delta_k [\mathbf{Z}_k - \bar{\mathbf{Z}}(T_i)]' [\mathbf{Z}_k - \bar{\mathbf{Z}}(T_i)]. \tag{8.4.5}$$

Así, el estimador de $\alpha = (\alpha_1, \ldots, \alpha_p)$ es

$$\widehat{\alpha} = \mathbf{A}^{-1} \mathbf{B}' \tag{8.4.6}$$

y el estimador de la varianza de $\widehat{\alpha}$ es

$$\widehat{\mathbf{V}} = \widehat{Var}(\widehat{\alpha}) = \mathbf{A}^{-1} \mathbf{C} \mathbf{A}^{-1}. \tag{8.4.7}$$

Para verificar la hipótesis $H_{0j} = 0$, $j = 1, \ldots, p$ se emplea la estadística

$$Z_j = \frac{\widehat{\alpha}_j}{\sqrt{\widehat{\mathbf{V}}_{jj}}}. \tag{8.4.8}$$

La estadística Z_j, bajo la hipótesis nula y para valores de n grandes, tiene distribución normal.

Para verificar la hipótesis sobre un subconjunto de parámetros contenidos en α, $H_{0j}: \alpha_j = \alpha_{0j}$ para todo $j \in \mathbf{J}$, es decir, α_{0j} es un subvector de α. La estadística es:

$$\chi^2 = [\widehat{\alpha}_j - \alpha_{0j}]' \widehat{\mathbf{V}}_j^{-1} [\widehat{\alpha}_j - \alpha_{0j}], \tag{8.4.9}$$

donde $\widehat{\mathbf{V}}_j$ es la matriz de las covarianzas correspondientes a las covariables asociadas con los parámetros α contenidos en α. Bajo la hipótesis nula, la estadística χ^2 tiene distribución ji-cuadrado con grados de libertad iguales al tamaño del subvector α_j.

Capítulo

nueve

Análisis bayesiano en datos de sobrevida

$$P(A|B)P(B) = P(A \cap B)$$

9.1. Introducción

En el análisis frecuentista de datos, se asume que la información sobre un parámetro θ fijo y desconocido está contenido en la muestra "actual" de la población definida por el parámetro. En el análisis estadístico bayesiano se asume lo anterior, además de la información preliminar, o *a priori*, sobre la distribución del parámetro. Así, el parámetro, θ, se considera como una variable aleatoria.

Para el análisis de tiempo hasta la ocurrencia de un evento, el investigador, basándose en la experiencia previa del proceso en observación o en el concepto de otros investigadores en el área, expresa esta información preliminar en la función de sobrevida. La información de la muestra está contenida en la función de verosimilitud. De esta manera, la información disponible sobre el parámetro, a partir de la perspectiva bayesiana, consta de dos componentes: la información preliminar, histórica o *a priori*, y la información contenida en la muestra "actual". Estos dos componentes de información sobre θ son considerados e incluidos en el teorema de Bayes, con el cual se obtiene una distribución *a posteriori* (posmuestra) de la distribución de la función sobrevida, la cual es la distribución de la función de sobrevida dados los datos.

En el análisis bayesiano, los parámetros del modelo se consideran como variables aleatorias seleccionadas desde una distribución *a priori*. Usualmente, la distribución *a priori* contiene y refleja el conocimiento e incertidumbre que tiene el investigador sobre el parámetro. Se puede considerar la varianza de la distribución *a priori* como inversamente proporcional a la cantidad de información muestral. Este capítulo se orienta por los libros de Ibrahim, Chen y Sinha [34] y de Kaeding [36].

En el análisis de tiempo hasta la ocurrencia de un evento o el análisis de sobrevida, el interés es la función de sobrevida o, su equivalente, la función de riesgo acumulada. Esta se trata como una cantidad aleatoria muestreada desde algún proceso estocástico. Entonces, la información contenida en datos corresponde a una muestra de una población definida por la función de sobrevida, la cual se combinará con la información *a priori* para obtener la distribución de la función de sobrevida (*a posteriori*), dados los datos.

Así, en el análisis bayesiano, los parámetros del modelo se tratan como variables aleatorias y la inferencia sobre los parámetros se basa en la distribución *a posteriori* de los parámetros, dados los datos. La distribución *a posteriori* se obtiene usando el teorema de Bayes como la función de verosimilitud de los datos ponderados con una distribución previa. La distribución previa permite incorporar conocimiento o experiencia del rango probable

de valores de los parámetros de interés en el análisis. Cuando no se tiene conocimiento previo de los valores de los parámetros, se puede usar una distribución *a priori* no informativa, y los resultados del análisis bayesiano serán muy similares a los de un análisis clásico basado en la función de verosimilitud.

En este capítulo se resumen algunos elementos básicos del análisis bayesiano para el tratamiento de datos de sobrevida, se esquematiza el muestreador de Gibbs, se desarrollan algunos modelos de regresión paramétricos y algunos modelos de regresión semiparamétricos y, finalmente, se presenta una generalización del modelo de Cox. El tratamiento computacional se desarrolla mediante el paquete R y el paquete SAS.

9.2. Elementos del análisis estadístico bayesiano

Suponga que $\mathbb{T} = (t_1, \ldots, t_n)'$ es un vector de n observaciones de la variable aleatoria $\mathbf{T}$, cuya distribución de probabilidad $p(\mathbb{T}|\theta)$ depende de los valores de k parámetros $\theta = (\theta_1, \ldots, \theta_p)'$. Además, suponga que θ tiene una distribución de probabilidad $p(\theta)$. Entonces,

$$p(\mathbb{T}|\theta)p(\theta) = p(\mathbb{T}, \theta) = p(\theta|\mathbb{T})p(\mathbb{T}). \tag{9.2.1}$$

Dados los datos observados $\mathbb{T}$, la distribución condicional de θ es

$$p(\theta|\mathbb{T}) = \frac{p(\mathbb{T}|\theta p(\theta))}{p(\mathbb{T})}. \tag{9.2.2}$$

El paradigma bayesiano se basa en un modelo específico de probabilidad para los datos de tiempo hasta el evento, $\mathbb{T}$, dado un vector de parámetros desconocidos θ, lo cual conduce a la función de verosimilitud $L(\theta|\mathbb{T})$. Entonces, se asume que θ es aleatorio y tiene una distribución *a priori* denotada por $p(\theta)$. La inferencia en torno a θ se basa en la distribución *a posteriori*, la cual es obtenida por el teorema de Bayes. La distribución *a posteriori* de θ es dada por

$$p(\theta|\mathbb{T}) = \frac{p(\mathbb{T}|\theta)p(\theta)}{\int_\Theta p(\mathbb{T}|\theta)p(\theta)d\theta}. \tag{9.2.3}$$

Una escritura alternativa de (9.2.3) es

$$p(\theta|\mathbb{T}) = cp(\mathbb{T}|\theta)p(\theta), \tag{9.2.4}$$

donde $c^{-1} = \int_{\Theta} p(\mathbb{T}|\theta)p(\theta)d\theta$ es una constante de normalización, puesto que $p(\theta|\mathbb{T})$ es una distribución de probabilidad que debe integrar a 1. La constante c se denomina distribución *marginal* de los datos o distribución *a priori* predictiva.

Uno de los métodos de cómputo más populares para obtener muestras a partir de $p(\theta|\mathbf{T})$ es el llamado *muestreador de Gibbs*. Este es un algoritmo de simulación muy potente que permite obtener muestras desde $p(\theta|\mathbf{T})$ sin necesidad de conocer la constante de normalización c, como se afirma en Ibrahim, Chen y Sinha [34, pp. 18-22].

Ahora, dados los datos $\mathbb{T}$, $p(\theta|\mathbb{T})$ en (9.2.3) se puede considerar como una función de θ. Con esta consideración, la función de verosimilitud de θ, dado los datos $\mathbb{T}$, se escribe como $L(\theta|\mathbb{T})$. Así, la fórmula de Bayes se escribe como

$$p(\theta|\mathbb{T}) = L(\theta|\mathbb{T})p(\theta). \tag{9.2.5}$$

En otras palabras, el teorema de Bayes dice que la distribución de probabilidad para θ posterior a los datos $\mathbb{T}$ es proporcional al producto de la distribución *a priori* para θ y la verosimilitud para θ dado $\mathbb{T}$; aún más, el conocimiento que se pueda tener sobre θ es el resultado del conocimiento anterior actualizado con el conocimiento presente sobre θ. Esto es:

$$\text{Distribución } \textit{a posteriori} \propto \text{Verosimilitud} \times \text{Distribución a priori.}$$

La verosimilitud es la función a través de la cual los datos $\mathbb{T}$ modifican (actualizan) el conocimiento *a priori* de θ. En consecuencia, se puede considerar que la verosimilitud representa la información sobre θ procedente de los datos actuales.

Una ventaja de la expresión (9.2.5) es que provee una formulación matemática de cómo el conocimiento previo sobre θ puede combinarse con el conocimiento nuevo. En efecto, el teorema de Bayes permite actualizar información sobre el conjunto de parámetros contenidos en θ, en la medida en que se acopie información.

A continuación, se ilustra la afirmación anterior. Suponga que se tiene una muestra inicial de observaciones $\mathbb{T}_1$. Entonces, por la fórmula de Bayes, se tiene

$$p(\theta|\mathbb{T}_1) \propto p(\theta)L(\theta|\mathbb{T}_1). \tag{9.2.6}$$

Ahora, suponga que se tiene una segunda muestra de observaciones $\mathbb{T}_2$ distribuidas independientemente de la primera muestra, entonces,

$$\begin{aligned} p(\theta|\mathbb{T}_1, \mathbb{T}_2) &\propto p(\theta)L(\theta|\mathbb{T}_1)L(\theta|\mathbb{T}_2) \\ &\propto p(\theta|\mathbb{T}_1)L(\theta|\mathbb{T}_2). \end{aligned} \tag{9.2.7}$$

La expresión (9.2.7) es similar a la expresión (9.2.6), excepto que $p(\theta|\mathbb{T}_1)$, la distribución *a posteriori* para θ dado $\mathbb{T}_1$, hace el papel de la distribución *a priori* para la segunda muestra. Naturalmente, el proceso es inductible a cualquier numero de k-muestras secuenciales.

En resumen, "el teorema de Bayes describe el proceso de aprendizaje desde la experiencia, y muestra cómo el conocimiento sobre el estado de la naturaleza representado por θ es continuamente modificado y actualizado en la medida que se obtiene información nueva" [4, p. 11].

Otro aspecto importante del análisis bayesiano es la predicción, una tarea importante en los problemas de regresión. La *distribución a posteriori predictiva* de un vector de observaciones t, dado los datos $\mathbb{T}$, es definida como

$$p(t|\mathbb{T}) = \int_{\Theta} f(t|\theta)p(\theta|\mathbb{T})d\theta, \tag{9.2.8}$$

donde $f(t|\theta)$ denota la densidad de t y $p(\theta|\mathbb{T})$ es la distribución *a posteriori* de θ. Se observa que (9.2.8) es el valor esperado *a posteriori* de $f(t|\theta)$, por lo que el muestreo es fácil de obtener desde (9.2.8) mediante el muestreador de Gibbs de $p(\theta|\mathbb{T})$.

Esta es una de las buenas características del análisis bayesiano, ya que a partir de (9.2.8) se muestra que las predicciones y la distribución de los predictores son fáciles de computar una vez estén disponibles las muestras desde $p(\theta|\mathbb{T})$.

Toda la información de los datos, que proceden de la base de una distribución *a priori* y una muestra, está contenida en la distribución *a posteriori*. Una forma de resumir parcialmente la información contenida en la distribución *a posteriori* es acotar uno o más intervalos, los cuales contienen cantidades preestablecidas de probabilidad. Una situación diferente ocurre cuando no hay límites de interés especial, sino que es necesario mostrar un rango dentro del cual se encuentre la mayor parte de la distribución. Uno de tales intervalos debería tener la propiedad de que la densidad, para cada punto dentro del intervalo, sea más grande que para cada punto fuera del intervalo. Además, es deseable que para una probabilidad dada, el intervalo que la contiene sea lo más angosto posible. Dado que para este tipo de intervalos todo punto incluido tiene más alta densidad de probabilidad que cualquier punto excluido, se denominan estos intervalos como *de alta densidad a posteriori* o HPD, por sus siglas en inglés (*Highest Posterior Density*).

Una clase importante de distribuciones *a priori* son las *distribuciones conjugadas*. En las expresiones (9.2.4) y (9.2.5) para la *distribución a posteriori* se tienen tres distribuciones: (i) la *a posteriori* misma, $p(\theta|\mathbf{T})$, (ii) la distribución muestral (de los datos), $p(\mathbb{T}|\theta)$ y (iii) la distribución *a priori*, $p(\theta)$ de θ.

Suponga que la distribución de la muestra, $p(\theta|\mathbf{T})$, pertenece a una familia de distribuciones $\mathcal{F}$ y que la distribución *a priori* de θ, $p(\theta)$, pertenece a la familia de distribuciones $\mathcal{P}$. Se dice que $\mathcal{P}$ es *conjugada* con $\mathcal{F}$ si para toda $p(\cdot|\theta) \in \mathcal{F}$ y $p(\cdot) \in \mathcal{P}$, entonces, $p(\theta|\mathbf{T}) \in \mathcal{P}$.

La propiedad de "conjugación" se puede entender, intuitivamente, como una propiedad en la cual una distribución *a posteriori* "hereda" la distribución de la *a priori* a través de la distribución muestral.

9.3. Muestreo desde la distribución *a posteriori*

El muestreador de Gibbs es uno de los algoritmos de muestreo más conocidos dentro de los métodos tipo MCMC, acrónimo de *Markov Chain Monte Carlo*. Las cadenas de Markov son procesos que describen trayectorias en las que las cantidades sucesivas se describen probabilísticamente de acuerdo con el valor de sus predecesoras inmediatas. En muchos casos, estos procesos tienden a un equilibrio y las cantidades límite siguen una distribución invariante. Las técnicas de MCMC permiten la simulación a partir de una distribución "encriptada", que actúa como una distribución límite de una cadena de Markov y simula desde la cadena hasta que se alcanza el equilibrio (convergencia). La simulación de un solo valor por MCMC consiste en una secuencia completa de los valores de una cadena hasta que alcanza el equilibrio, y solo el valor de equilibrio se puede tomar como un valor simulado de la distribución límite, como se resume en el libro de Gamerman [28, pp. 4-5].

A continuación, se hace un resumen del algoritmo de Gibbs.

Considere el vector de parámetros $\theta = (\theta_1, \ldots, \theta_p)'$ y sea $p(\theta|\mathbb{T})$ su distribución *a posteriori*. El esquema básico del muestreador de Gibbs es como se describe a continuación:

Paso 0. Escoja un punto de inicio arbitrario $\theta_0 = (\theta_{01}, \ldots, \theta_{0p})'$ y haga $i = 0$.

Paso 1. Genere $\theta_{i+1} = (\theta_{i+1,1}, \ldots, \theta_{i+1,p})'$ de la siguiente manera:

- Genere $\theta_{i+1,1} \sim p(\theta_1|\theta_{i,2}, \ldots, \theta_{i,p}; \mathbb{T})$
- Genere $\theta_{i+1,2} \sim p(\theta_2|\theta_{i+1,1}, \theta_{i,3}, \ldots, \theta_{i,p}; \mathbb{T})$

  $\ldots \ldots \ldots$
- Genere $\theta_{i+1,p} \sim p(\theta_p|\theta_{i+1,1}, \theta_{i+1,2}, \ldots, \theta_{i+1,p-1}; \mathbb{T})$.

Paso 2. Haga $i = i + 1$, y vaya al paso 1.

Bajo ciertas condiciones de regularidad, el vector $\{\theta_i,\ i = 1, \ldots, \}$ tiene una distribución estacionaria[1] $p(\theta|\mathbb{T})$.

El algoritmo *Metropolis-Hastings* se resume a continuación.

Sea $\mathcal{U}(0, 1)$ una distribución uniforme sobre el intervalo $(0, 1)$. Entonces, una versión general del algoritmo de Metropolis-Hastings para muestrear a partir de la distribución a posterior $p(\theta|\mathbb{T})$ es como se reseña enseguida:

Paso 0. Escoja un punto inicial θ_0 y haga $i = 0$.

Paso 1. Genere un punto candidato θ^* desde $q(\theta_i, \cdot)$ y u desde $\mathcal{U}$.

- Sea $\theta_{i+1} = \theta^*$ si $u \leq a(\theta_i, \theta^*)$ y $\theta_{i+1} = \theta_i$ en otra parte, donde la probabilidad de aceptación es dada por

$$a(\theta, \vartheta) = \text{mín}\left\{\frac{p(\vartheta|\mathbb{T})q(\vartheta, \theta)}{p(\theta|\mathbb{T})q(\theta, \vartheta)}, 1\right\}, \tag{9.3.1}$$

  donde $q(\theta, \vartheta)$ es una densidad propuesta, la cual se considera como candidata a ser generadora de densidad, tal que

$$\int q(\theta, \vartheta)d\vartheta = 1.$$

Paso 3. Haga $i = i + 1$, y vaya al paso 1.

Es evidente que el desempeño del algoritmo de Metropolis-Hastings depende de la escogencia de la densidad $q(\cdot)$.

9.4. Modelos de regresión paramétricos

En la práctica, el análisis bayesiano se desarrolla sobre modelos paramétricos. En esta sección se esquematiza el análisis bayesiano para algunos modelos paramétricos del análisis de sobrevida. Un tratamiento más amplio se encuentra en Ibrahim, Chen y Sinha [34, cap. 2].

9.4.0.1. Modelo exponencial

El modelo exponencial en análisis de sobrevida es como el modelo normal en la mayoría de los métodos estadísticos. Suponga que se tienen n tiempos al evento, $\mathbf{t} = (t_1, \ldots, t_n)'$, independientes e idénticamente distribuidos

[1] Esto es, si las distribución conjunta de $(\theta_1, \ldots, \theta_p)$ y de $(\theta_{1+s}, \ldots, \theta_{p+s})$ es la misma para cualquier valor de s.

(*iid*), cada uno de los cuales tiene una distribución exponencial con parámetro θ. Se nota el indicador de censura δ_i, donde $\delta_i = 0$ si t_i es censurado y $\delta_i = 1$ si t_i es el tiempo al evento. Sea $f(t_i|\theta) = \theta \exp\{-\theta t_i\}$ la función de densidad para t_i, por lo que $S(t_i|\theta) = \exp\{-\theta t_i\}$ es la función de sobrevida y $\mathbb{T} = (n, t_i, \delta_i)$ denota los datos observados. La función de verosimilitud de θ es

$$L(\theta|\mathbb{T}) = \prod_{i=1}^{n} f(t_i\theta)^{\delta_i} S(t_i|\theta)^{(1-\delta_i)} = \theta^d \exp\left(-\theta \sum_{i=1}^{n} t_i\right), \tag{9.4.1}$$

donde $d = \sum_{i=1} n\delta_i$. La conjugada *a priori* para θ es la *a priori* gama. Sea $\mathcal{G}(\alpha_0, \theta_0)$, con densidad dada por

$$p(\theta|\alpha_0, \theta_0) \propto \theta^{\alpha_0 - 1} \exp\{-\theta_0\theta\}.$$

Entonces, tomando una $\mathcal{G}(\alpha_0, \theta_0)$ como *a priori* de θ, el parámetro de la distribución de θ es dado por

$$\begin{aligned} p(\theta|\mathbb{T}) &\propto L(\theta|\mathbb{T})p(\theta|\alpha_0, \theta_0) \\ &\propto \left(\theta^{\sum_{i=1}^{n} \delta_i} \exp\left\{-\theta \sum_{i=1}^{n} t_i\right\}\right)(\theta^{\alpha_0 - 1} \exp(-\theta_0\theta)) \\ &= \theta^{\alpha_0 + d - 1} \exp\left\{-\theta(\theta_0 + \sum_{i=1}^{n} t_i)\right\}. \end{aligned} \tag{9.4.2}$$

Así, se reconoce el núcleo de la distribución *a posteriori* en (9.4.2) como una distribución $\mathcal{G}(\alpha_0 + d, \theta_0 + \sum_{i=1}^{n} t_i)$. La media y varianza *a posteriori* de θ son dadas por

$$E(\theta|\mathbb{T}) = \frac{\alpha_0 + d}{(\theta_0 + \sum_{i=1}^{n} t_i)}$$

y

$$Var(\theta|\mathbb{T}) = \frac{\alpha_0 + d}{(\theta_0 + \sum_{i=1}^{n} t_i)^2}.$$

La distribución predictiva *a posteriori* de un valor futuro de tiempo t_f, teniendo en cuenta la definición de la función gama (3.7.2), es dado por

$$\begin{aligned} p(t_f|\mathbb{T}) &= \int_0^\infty p(t_f|\theta)p(\theta|\mathbb{T})d\theta \\ &\propto \int_0^\infty \theta^{\alpha_0 + d + 1 - 1} \exp\left\{-\theta(t_f + \theta_0 + \sum_{i=1}^{n} t_i)\right\} d\theta. \\ &= \Gamma(\alpha_0 + d + 1)\left(\alpha_0 + \sum_{i=1}^{n} t_i + t_f\right)^{-(d + \alpha_0 + 1)} \end{aligned}$$

La distribución predictiva *a posteriori* normalizada está dada por

$$p(t_f|\mathbb{T}) = \begin{cases} \frac{(d+\alpha_0)(\theta_0+\sum_{i=1}^n t_i)^{(\theta_0+d)}}{(\theta_0+\sum_{i=1}^n t_i+t_f)^{(\theta_0+d+1)}}, & si\ t_f > 0 \\ 0, & \text{en otro caso.} \end{cases} \tag{9.4.3}$$

Esta distribución predictiva es conocida como una distribución *inversa beta*.

Para construir un *modelo de regresión*, se introducen covariables a través del parámetro θ. Se escribe

$$\theta(\mathbf{x}_i) = g(\mathbf{x}_i'\beta) = g(\beta_1 x_1 + \cdots + \beta_p x_p), \tag{9.4.4}$$

donde $\mathbf{x}_i$ es un vector de covariables de tamaño $p \times 1$, β es un vector de tamaño $p \times 1$ de coeficientes de regresión, y $g(\cdot)$ es una función conocida. Una forma de g es $g(\mathbf{x}_i'\beta) = \exp\{\mathbf{x}_i'\beta\}$, y otra forma es $g(\mathbf{x}_i'\beta) = (\mathbf{x}_i'\beta)^{-1}$. Se puede optar por otras funciones del predictor $\mathbf{x}_i'\beta$ de acuerdo con los datos disponibles.

En este texto se emplea el siguiente modelo de regresión paramétrico:

$$\theta(\mathbf{x}_i) = g(\mathbf{x}_i'\beta) = \exp\{\beta_1 x_1 + \cdots + \beta_p x_p\}. \tag{9.4.5}$$

La correspondiente función de verosimilitud es

$$\begin{aligned} L(\beta|\mathbb{T}) &= \prod_{i=1}^n f(t_i|\theta_i)^{\delta_i} S(t_i|\theta_i)^{1-\delta_i} \\ &= \left[\exp\{\mathbf{x}_i'\beta\}\exp(-t_i\exp\{\mathbf{x}_i'\beta\})\right]^{\delta_i}\left[\exp(-t_i\exp(\mathbf{x}_i'\beta))\right]^{1-\delta_i} \\ &= \exp\left\{\sum_{i=1}^n \delta_i \mathbf{x}_i'\beta\right\}\exp\left\{-\sum_{i=1}^n t_i\exp(\mathbf{x}_i'\beta)\right\}. \end{aligned} \tag{9.4.6}$$

Para la verosimilitud expresada en (9.4.6), se define el conjunto de datos como $\mathcal{D} = (n, t, \mathbb{X}, \delta)$, donde $\mathbb{X}$ es la matriz de tamaño $n \times p$ de covariables, con la i-ésima fila $\mathbf{x}_i'$.

Entre las distribuciones *a priori* para β se incluyen la distribución uniforme impropia, es decir, $p(\beta) \propto 1$, y la distribución normal.

Suponga que β tiene distribución normal p-variante, $\beta \sim N_p(\mu_0, \Sigma_0)$, donde μ_0 es la media *a priori* y Σ_0 es la matriz de covarianzas *a priori*. Así, la distribución *a posteriori* de β es dada por

$$p(\beta|\mathcal{D}) =\propto L(\beta|\mathcal{D})p(\beta|\mu_0, \Sigma_0), \tag{9.4.7}$$

donde $p(\beta|\mu_0, \Sigma_0)$ es la distribución *a priori* y corresponde a la distribución normal p-variante, $N_p(\mu_0, \Sigma_0)$.

La distribución *a posteriori* dada en (9.4.7) no tiene una forma cerrada, por lo que son necesarios métodos de tipo MCMC para muestrear a partir de la distribución *a posteriori* de β. Para desarrollar el muestreador de Gibbs de este modelo, se debe escribir la distribución condicional completa. Sea β_j el j-ésimo componente de β y sea $\beta_{(-j)}$ el vector β sin el j-ésimo elemento. Entonces, la j-ésima distribución condicional completa puede escribirse como

$$p(\beta_j|\mathcal{D}) =\propto L(\beta_j, \beta_{(-j)}|\mathcal{D})p(\beta_j, \beta_{(-j)}|\mu_0, \Sigma_0), \text{ con } j = 1, 2, \ldots, p. \tag{9.4.8}$$

Para muestrear a partir de cada una de las distribuciones condicionales completas contenidas en (9.4.8), se debe emplear un algoritmo de aceptación-rechazo. Para esto se debe demostrar la concavidad o la log-concavidad de la respectiva función, en Ibrahim, Chen y Sinha [34, p. 33] se hace tal demostración.

Dentro de los resultados esperados sobre la distribución, está la función de sobrevida evaluada en un valor particular del tiempo t. Para el modelo de regresión exponencial, la función de sobrevida de una unidad con vector de covariables $\mathbf{x}$ en un punto t es dado por $S(t) = \exp\{-\exp(\mathbf{x}'\beta)t\}$. De manera que si se obtienen muestras desde la distribución *a posteriori* de β, se pueden calcular resúmenes *a posteriori* de $S(t)$, tal como la media *a posteriori*, la varianza y los intervalos de credibilidad.

Ejemplo 9.4.1. Se sigue una muestra de 38 personas y se registra el tiempo hasta el alivio de un dolor de cabeza. El conjunto de datos contiene la variable *minutos*, que representa el tiempo reportado para el alivio del dolor de cabeza, la variable *grupo* al que está asignado el paciente (grupo 1, 0, 2) y la variable *censura*, que es una variable binaria que indica si la observación del tiempo hasta el alivio está censurada o no; es decir, si hay alivio (*censura* = 1) o si no lo hay (*censura* = 0). Datos tomados de SAS/STAT [62, p. 5002]

Con el procedimiento LIFEREG del paquete SAS, se ajusta el modelo para una sola covariable, *grupo*, por lo tanto, $\beta = (\beta_0, \beta_1)'$, donde β_0 denota el término del intercepto y β_1 es el coeficiente de la covariable grupo. El siguiente es el código SAS para ajustar el modelo.

```
data dolor_ca;
input Minutos Grupo Censura @@;
datalines;
11 1 0 12 1 0 19 1 0 19 1 0
19 1 0 19 1 0 21 1 0 20 1 0
21 1 0 21 1 0 20 1 0 21 1 0
```

```
20 1 0 21 1 0 25 1 0 27 1 0
30 1 0 21 1 1 24 1 1 14 2 0
16 2 0 16 2 0 21 2 0 21 2 0
23 2 0 23 2 0 23 2 0 23 2 0
25 2 1 23 2 0 24 2 0 24 2 0
26 2 1 32 2 1 30 2 1 30 2 0
32 2 1 20 2 1
;
proc lifereg data=dolor_ca;
model Minutes*Censor(1)=Group / dist=exponential;
bayes seed=1 outpost=Post;
run;
ods graphics off;
```

Por defecto, se asume *a priori* una distribución uniforme para el coeficiente de intersección. La distribución uniforme tomada como *a priori* anteriormente es plana. La recta real tiene una distribución que refleja el desconocimiento acerca de la ubicación del parámetro. Esta distribución asigna la misma probabilidad a todos los valores posibles que este coeficiente de regresión pueda tomar. Dado que se asume una distribución *a priori* no informativa, los resultados coinciden con los obtenidos mediante la máxima verosimilitud.

La tabla 9.1 contiene los resultados de la media, la desviación estándar y los percentiles posteriores (cuartiles) de los parámetros, con el uso de $N = 10000$ muestras Gibbs (cadena). Las figuras 9.1 y 9.2 muestran la

Tabla 9.1. Resumen posterior de β con el modelo exponencial

Datos del dolor de cabeza

Parámetro	N	Media	Desviación estándar	Percentiles		
				25 %	50 %	75 %
Intercepto	10000	27.460	0.5676	23.545	27.346	31.170
Grup0	10000	0.4152	0.3769	0.1629	0.4128	0.6660

trayectoria o trazabilidad (panel superior) muestreador de Gibbs y las densidades marginales *a posteriori* de β_0 y β_1 usando el modelo exponencial (Weibull, con parámetro de escala igual a 1). La trazabilidad indica que el

muestreador de Gibbs maximiza adecuadamente. Se observa que la media de los parámetros se aproxima a la moda de cada uno de estos.

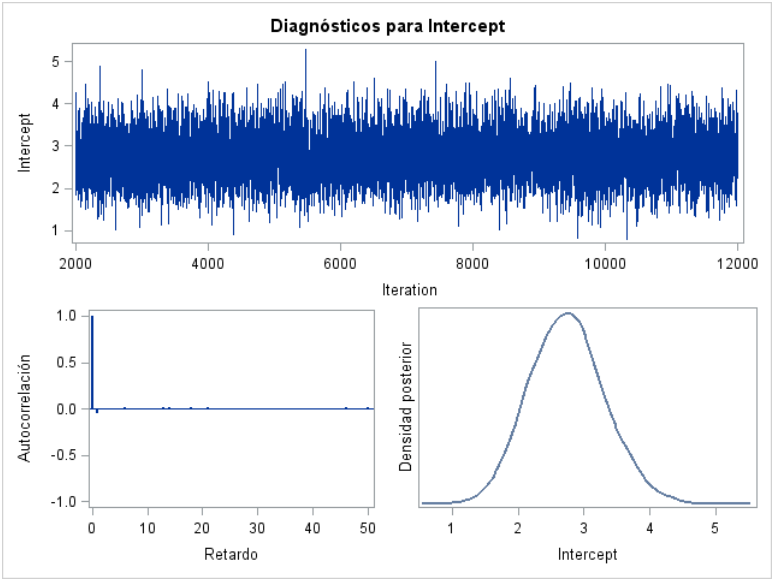

Figura 9.1. Trazabilidad del Gibbs y densidad marginal posterior de β_0.

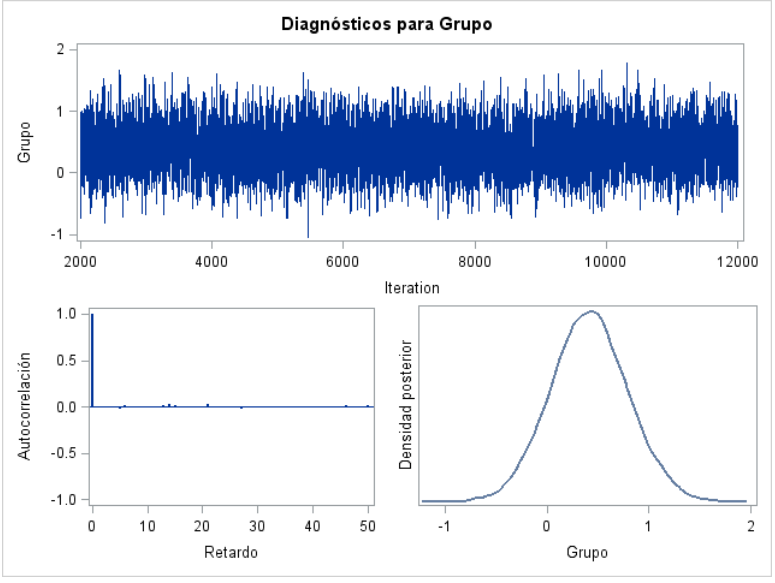

Figura 9.2. Trazabilidad del Gibbs y densidad marginal posterior de β_1.

El intervalo de credibilidad del 95 % para β_0 es (1.6977, 3.9026), mientras que para β_1 es (−0.3179, 1.1573). En la distribución *a posteriori* de β_1, gran parte de la masa está a la derecha de 0, por lo que, de acuerdo con la definición de cada grupo, se hablaría a favor de uno de ellos.

9.4.0.2. Modelo Weibull

Este es tal vez el modelo de mayor uso entre los modelos de sobrevida paramétricos. Suponga que se tienen n tiempos al evento, $\mathbf{t} = (t_1, \ldots, t_n)'$, independientes e idénticamente distribuidos (*iid*), cada uno de los cuales tiene una distribución $Weibull(\theta, \alpha)$ tal como se definió en (3.4.1). Es más conveniente escribir este modelo en términos de la reparametrización $\lambda = \ln \alpha$, con la cual se tiene la función de probabilidad

$$f(t|\theta, \lambda) = \theta t^{\theta-1} \exp\{\lambda - \exp(\lambda) t^{\theta}\}. \tag{9.4.9}$$

La función de sobrevida asociada con la distribución $Weibull(\theta, \lambda)$ es

$$S(t|\theta, \lambda) = \exp\{-\exp(\lambda) t^{\theta}\}.$$

La función de verosimilitud de (θ, λ), dado el conjunto de datos estructurado en $\mathcal{D} = (n, t, \mathbb{X}, \delta)$ definido para el caso exponencial, es

$$L(\theta, \lambda|\mathcal{D}) = \prod_{i=1}^{n} f(t|\theta, \lambda)^{\delta_i} S(t|\theta, \lambda)^{1-\delta_i}$$
$$\theta^d \exp\left\{ d\lambda + \sum_{i=1}^{n} (\delta_i(\theta - 1)\ln(t_i) - \exp(\lambda) t_i^{\theta}) \right\}. \tag{9.4.10}$$

Cuando se asume que θ es conocido, la distribución *a priori* conjugada para $\exp(\lambda)$ es la *a priori* gama. Ninguna conjunta *a priori* conjugada está disponible cuando los dos parámetros (θ, λ) son desconocidos. En este caso, una especificación de la *a priori* conjunta es considerar a θ y λ independientes, donde θ tiene una distribución gama y λ tiene una distribución normal. Sea $\mathcal{G}(\theta_0, \kappa_0)$ la distribución *a priori* gama para θ, y $n(\mu_0, \sigma_0^2)$ la distribución *a priori* normal para λ. La distribución conjunta *a posteriori* para (θ, λ) está

dada por

$$p(\theta, \lambda|\mathcal{D}) \propto L(\theta, \lambda|\mathcal{D})p(\theta|\theta_0, \kappa_0)p(\lambda|\mu_0, \sigma_0^2)$$
$$\propto \left[\prod_{i=1}^{n} f(t|\theta, \lambda)^{\delta_i} S(t|\theta, \lambda)^{1-\delta_i}\right] p(\theta|\theta_0, \kappa_0)p(\lambda|\mu_0, \sigma_0^2)$$
$$= \theta^{\theta_0+d-1} \exp\left\{d\lambda + + \sum_{i=1}^{n}(\delta_i(\theta-1)\ln(t_i) - \exp(\lambda)t_i^{\theta})\right.$$
$$\left. -\kappa_0\theta - \frac{1}{2\sigma_0^2}(\lambda-\mu_0)^2\right\}. \quad (9.4.11)$$

La distribución conjunta *a posteriori* de (θ, λ) no tiene una forma cerrada, pero Ibrahim, Chen y Sinha [34] señalan que las condicionales *a posteriori* $(\theta|\lambda, \mathcal{D})$ y $(\lambda|\theta, \mathcal{D})$ son las log-concavas, con lo cual el muestreador de Gibbs es sencillo para este modelo.

Para construir *un modelo de regresión Weibull*, se introducen las covariables a través del parámetro λ, y se escribe $\lambda_i = \mathbf{x}_i'\beta$. La distribución *a priori* común para β incluye la uniforme impropia, es decir, $p(\beta) \propto 1$ y una normal. Se asume una $N_p(\mu_0, \Sigma_0)$ como la *a priori* para β y una gama como la *a priori* para θ. Se tiene así que la conjunta *a posteriori* es

$$p(\beta, \theta|\mathcal{D}) \propto \theta^{\alpha_0+d-1} \exp\left\{\sum_{i=1}^{n}(\delta_i\mathbf{x}_i'\beta + \delta_i(\theta-1)\ln(t_i) - t_i^{\theta}\exp(\mathbf{x}_i'\beta))\right.$$
$$\left. -\kappa_0\theta - \frac{1}{2}(\beta-\mu_0)'\Sigma_0^{-1}(\beta-\mu_0)\right\}. \quad (9.4.12)$$

Formas cerradas para la distribución *a posteriori* de β generalmente no están disponibles, por lo tanto, se requiere de integración numérica o de los métodos MCMC. En la mayoría de los paquetes estadísticos como R, SAS o BUGS, se puede ajustar fácilmente el modelo de regresión (9.4.12) empleando los métodos MCMC incorporados en estos.

Ejemplo 9.4.2. Considere los datos sobre pacientes con melanoma de un ensayo clínico descrito en Ibrahim, Chen y Sinha [34, pp. 4 y 36]. El tiempo hasta el evento muerte se modela mediante un modelo de regresión Weibull con tres covariables.

Hubo 285 pacientes en este estudio (referenciado como e1684) a quienes se les registraron las siguientes 5 variables:

- TRT: tratamiento, 0 = Grupo Control y 1 = Grupo IFN.
- TPO: tiempo observado libre de la enfermedad.

- CEN: indicador de censura, 1 = ocurre el eveno de interés, y 0 = censura.
- EDAD: edad centrada en la media.
- SEXO: 0 = Hombre, 1 = Mujer.

Una lista parcial de los datos se muestra para 10 pacientes (entre los 285) con el valor de las variables anteriores, editados para procesar en el paquete SAS:

```
ID  TRT   TPO   CENS    EDAD    SEXO
1    1  1.15068  1 -11.03594366   0
2    1  0.62466  1  -5.12904366   0
3    0  1.89863  0  23.18595634   1
4    0  0.45479  1  11.14485634   1
5    0  2.09041  1 -13.32084366   0
6    1  9.38356  0   0.94215634   0
7    0  1.69863  1 -15.20854366   1
9    0  0.25753  1  -6.31534366   1
10   0  9.64384  0 -14.08254366   0
```

El procesamiento de los datos de tiempo hasta quedar libre de la enfermedad (melanoma), con censura a la derecha, se realiza con el procedimiento *PROC LIFEREG* del paquete SAS, el cual produce las estimaciones bayesianas de los coeficientes de regresión mediante el uso del siguiente código:

```
ods graphics on;
proc lifereg data=e1684;
class SEXO;
model TPO*CEN(1)=EDAD SEXO TRT / dist=Weibull;
bayes WeibullShapePrior=gamma
```

Con esta sintaxis, se ajusta un modelo cuyas covariables son la edad, el sexo y el tratamiento. Se asume que la variable *tiempo hasta el evento* tiene una distribución Weibull. Además, el parámetro de forma de esta distribución sigue una distribución gama.

En la tabla 9.2 se muestran las estimaciones asociadas con los β's posteriores y el parámetro de forma de la Weibull (θ) para el modelo ajustado, incluyendo la media, la desviación estándar, dos cuartiles y la mediana.

Tabla 9.2. Resumen posterior de los β's con el modelo Weibull

Datos del melanoma E1684

Parámetro	N	Media	Desviación estándar	Percentiles 25 %	50 %	75 %
Intercept	10000	1.9084	0.0481	1.8755	1.9076	1.9401
EDAD	10000	-0.00330	0.00176	-0.00449	-0.00330	-0.00214
SEXO0	10000	0.0648	0.0492	0.0317	0.0647	0.0981
TRT	10000	0.0392	0.0508	0.00606	0.0402	0.0735
FormaWeib	10000	4.4480	0.3895	4.1852	4.4373	4.7073

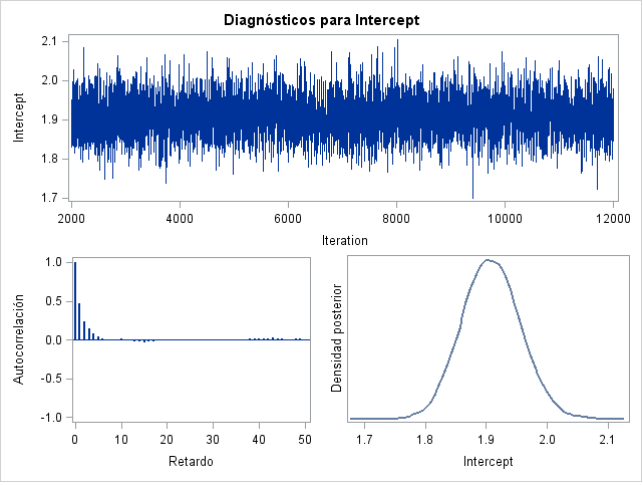

Figura 9.3. Trazabilidad del Gibbs y densidad marginal posterior de β_0.

Las figuras 9.3 a 9.7 contienen las gráficas asociadas con la trazabilidad del muestreador de Gibbs y las densidades marginales para los respectivos parámetros *a posteriori* del modelo ajustado. La trazabilidad del proceso de los Gibbs muestra una buena mezcla de estos para todos los parámetros.

Para otros modelos tales como los de valor extremo, log-normal y gama se puede consultar a Ibrahim, Chen y Sinha [34, pp. 37-42].

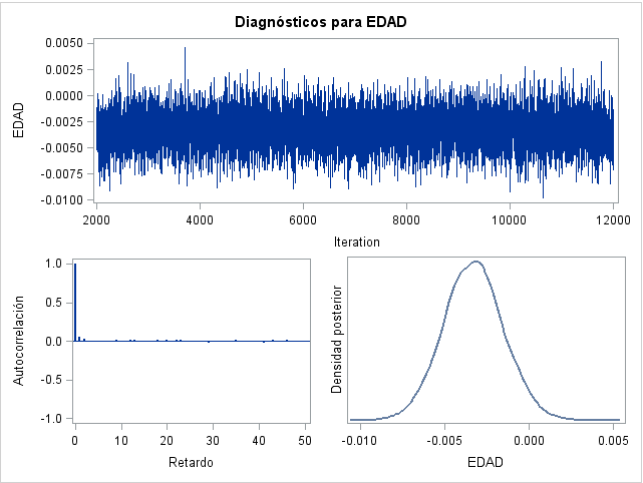

Figura 9.4. Trazabilidad del Gibbs y densidad marginal posterior de β_1.

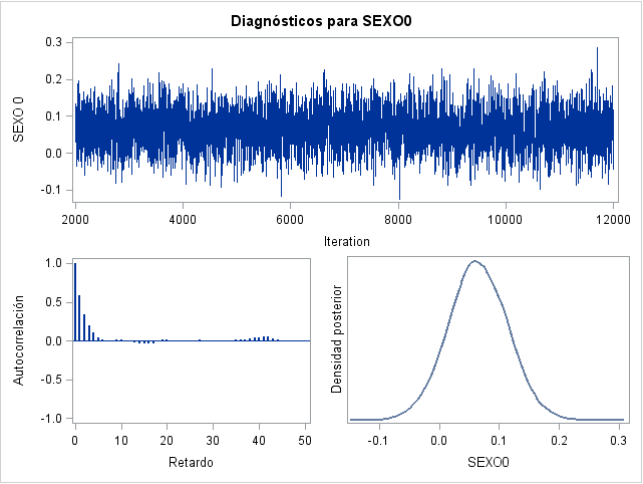

Figura 9.5. Trazabilidad del Gibbs y densidad marginal posterior de β_2.

9.5. Modelos de regresión semiparamétricos

En esta sección se hace una presentación del modelo de riesgos proporcionales de Cox. Se examinan los modelos de riesgo constante a trozos (en inglés, *piecewise*). Para cada caso, se desarrolla el proceso de la *a priori*, se construye la función de verosimilitud, se deriva la distribución *a posteriori* y se esquematiza la respectiva técnica MCMC para la inferencia.

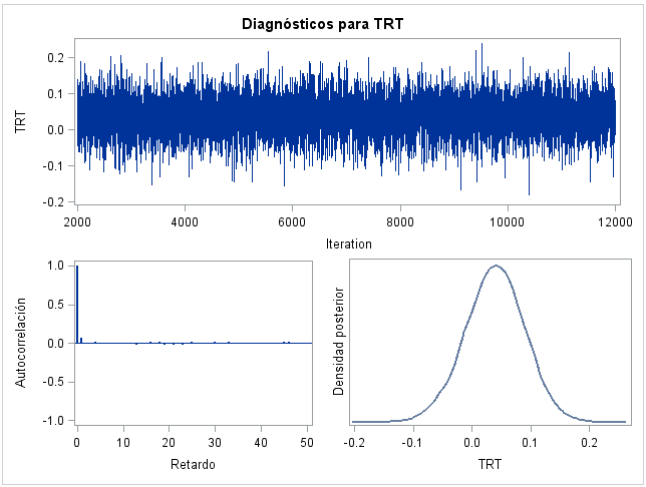

Figura 9.6. Trazabilidad del Gibbs y densidad marginal posterior de β_3.

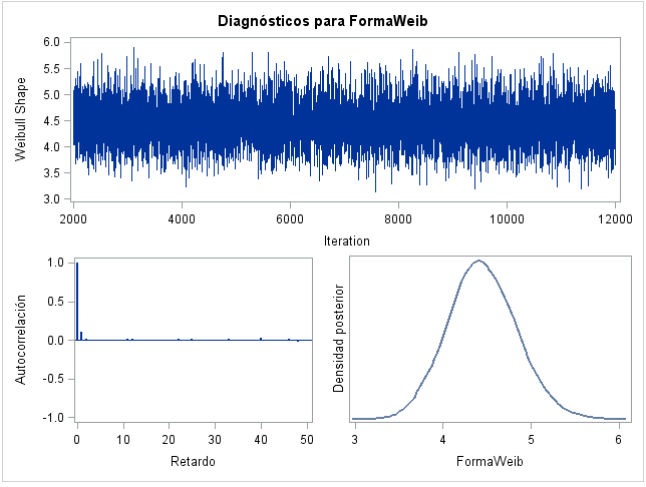

Figura 9.7. Trazabilidad del Gibbs y densidad marginal posterior de β_4.

9.5.0.1. Modelos de riesgo constante a trozos

Para la construcción de este modelo, se define una partición finita del eje de tiempo en la forma $0 < s_1 < \cdots < s_J$, con $s_J > s_i$, para todo $i = 1, \ldots, n$, Así, se construyen J intervalos $(0, s_1]$, $(s_1, s_2]$, $\ldots$, $(s_{J-1}, s_J]$. En el j-ésimo intervalo, se asume el riesgo base constante $h_0(t) = \lambda_j$ para $t \in I_j(s_{j-1}, s_j]$. En la figura 9.8 se muestra la función de riesgo base constante a trozos o tramos.

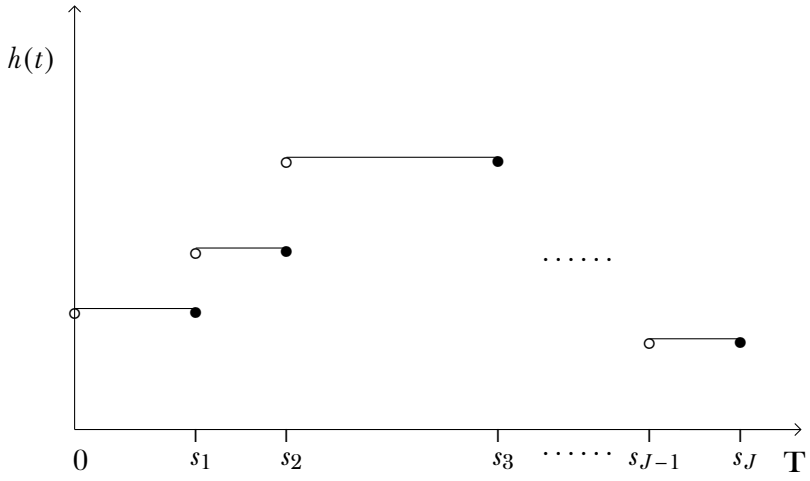

Figura 9.8. Función de riesgo base constante a trozos.

Sea $\mathcal{D} = (n, \mathbf{t}, \mathbb{X}, \delta)$ el arreglo que hace referencia a los datos observados, donde $\mathbf{t} = (t_1, \ldots, t_n)'$, $\delta = (\delta_1, \ldots, \delta_n)'$, con $\delta_i = 1$ si a la i-ésima unidad le ha ocurrido el evento y 0 en otro caso, y $\mathbb{X}$ es la matriz de tamaño $n \times p$ de covariables, con la i-ésima fila $\mathbf{x}_i'$ que define la i-ésima unidad. Se define $\lambda = (\lambda_1, \ldots, \lambda_J)'$, por lo que la función de verosimilitud de (β, λ) para las n-unidades es

$$L(\beta, \lambda|\mathcal{D}) = \prod_{i=1}^{n}\prod_{j=1}^{J}(\lambda_j \exp(\mathbf{x}_i'\beta))^{\nu_{ij}\delta_i} \times \exp\left\{-\nu_{ij}\left[\lambda_j(t_i - s_{j-1}) + \sum_{g=1}^{j-1}\lambda_g(s_g - s_{g-1})\right]\exp(\mathbf{x}_i'\beta)\right\}, \quad (9.5.1)$$

donde $\nu_{ij} = 1$, si a la i-ésima unidad le ocurre el evento o fue censurada en el j-ésimo intervalo, y es 0 en otro caso, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ es el vector $p \times 1$ de covariables para la i-ésima unidad, y $\beta = (\beta_1, \ldots, \beta_p)'$ es el correspondiente vector de coeficientes de regresión.

El modelo semiparamétrico en (9.5.1) se denomina como un *modelo exponencial a trozos* y puede acomodar varias formas de la función de riesgo base sobre los intervalos. Note que si $J = 1$, el modelo se reduce a un modelo paramétrico exponencial con tasa de falla $\lambda \equiv \lambda_1$.

Una distribución *a priori* del riesgo base λ es la gama independiente $\lambda_j \sim \mathcal{G}(\alpha_{0j}, \lambda_1 0j)$ para $j = 1, \ldots, J$. La verosimilitud dada en (9.5.1) se basa en los datos de sobrevida continuos. La función de verosimilitud basada en

datos agrupados o discretizados es dada por

$$L(\beta, \lambda | \mathcal{D}) \propto \prod_{j=1}^{J} G_j^*, \tag{9.5.2}$$

donde

$$G_j^* = \exp\left\{-\lambda_j \Delta_j \sum_{k \in \mathcal{R}_j - \mathcal{D}_j} \exp(\mathbf{x}_i'\beta)\right\} \times \prod_{l \in \mathcal{D}_j} [1 - \exp\{-\lambda_j \Delta_j \exp(\mathbf{x}_i'\beta)\}], \tag{9.5.3}$$

donde $\Delta_j = s_j - s_{j-1}$ es el ancho del j-ésimo intervalo, $\mathcal{R}_j$ es el conjunto de unidades en riesgo al evento, y $\mathcal{D}_j$ es el conjunto de unidades a las que les ocurrió el evento en el j-ésimo intervalo.

9.6. Una generalización del modelo de Cox

Se propone un modelo que puede ser considerado como una generalización del modelo de riesgo constante a trozos presentado en la sección anterior. Sin pérdida de generalidad, se asume una sola covariable.

Considere una función de riesgo constante a trozos $h(t|x) = \lambda_k \exp(x\beta_k)$, para $t \in I_k$, donde $I_k = (s_{k-1}, s_k]$ para $k = 1, \ldots, J$, y x univariada. Las siguientes especificaciones caracterizan el modelo:

(i) $\lambda_k \overset{indep.}{\sim} \mathcal{G}(\eta_k, \gamma_k)$, para $k = 1, \ldots, J$;

(ii) $\beta_{k+a}|\beta_0, \beta_1, \ldots, \beta_k \sim N(\beta_k, \omega_k^2)$ para $k = 0, \ldots, J-1$; y

(iii) los β_k's son independientes *a priori* de $\lambda = (\lambda_1, \ldots, \lambda_J)$.

Se asume que los hiperparámetros η_k, γ_k, ω_k y β_0 son conocidos. Este modelo es una versión discretizada de un modelo de riesgos no proporcionales junto con una versión discretizada del proceso gama *a priori* para el riesgo base $h_0(\cdot)$, donde η_k/γ_k es la media *a priori* y η_k/γ_k^2 es la varianza *a priori* del parámetro λ_k.

Los datos provenientes de estudios previos pueden emplearse para obtener la media *a priori* η_k/γ_k y la varianza *a priori* η_k/γ_k^2 de λ_k. Otra alternativa es seleccionar hiperparámetros para representar opiniones que son cercanamente no informativas en el subconjunto del espacio de parámetros soportado en la verosimilitud.

Los datos observados desde n unidades se escriben como el conjunto $\mathcal{D}_{obs} = \{X - i, (a_{li}, a_{ri}],\ i = 1, \ldots, n\}$, donde el tiempo al evento t_i para la i-ésima unidad se sabe que está dentro del intervalo $(a_{li}, a_{ri}]$, con $a_{li} < a_{ri}$ siendo dos de los puntos de grilla $(a_1, a_2, \ldots, a_J)$, que no son necesariamente consecutivos. Además, x_i es la covariable para la i-ésima unidad.

Sea $\mathcal{D} = \{t_i, x_i; i =, \ldots, n\}$ los datos completos, y $(\beta', \lambda')'$, donde $\beta = (\beta_1, \ldots, \beta_J)$.

La distribución de $(t_i|x_i, \beta', \lambda')$ es una exponencial a trozos. Así, la verosimilitud para los datos completos es dada por

$$L(\beta, \lambda|\mathcal{D}) \propto \prod_{k=1}^{J} \left[\lambda_k^{d_k} \exp\left\{ -\lambda_k \sum_{j \in \mathcal{R}_k} \exp(x_j \beta_k) \Delta_{jk} \right\} \exp\left(\sum_{j \in \mathcal{D}_k} x_j \beta_k \right) \right], \tag{9.6.1}$$

donde $\mathcal{R}_k$ es el conjunto de unidades en riesgo en a_{k-1}, $\Delta_{jk} = \text{mín}(t_j, a_k) - a_{k-1}$, $\mathcal{D}_k$ es el conjunto de unidades a las cuales les ocurre el evento en el intervalo $I_k = (a_{k-1}, a_k]$, y d_k es el número de unidades en $\mathcal{D}_k$.

Se emplea el muestreador de Gibbs para muestrear a partir de la distribución *a posteriori* de $(\beta', \lambda')'$. Para las condicionales completas, se obtiene

$$(\lambda_k|\lambda^{(-k)}, \beta, \mathcal{D}) \sim \mathcal{G}\left(\eta_k + d_k; \gamma_k + \sum_{j \in \mathcal{R}_k} \exp(x_j \beta_k) \Delta_{jk} \right), \tag{9.6.2}$$

$$p(\beta_k|\beta^{(-k)}, \lambda, \mathcal{D}) \propto \phi(\beta_k|\mu_k, \sigma_k^2) \left(-\lambda_k \sum_{j \in \mathcal{R}_k} \exp(x_j \beta_k) \Delta_{jk} \right), \tag{9.6.3}$$

donde $\phi(\beta_k|\mu_k, \sigma^2)$ es la densidad normal univariada $n(\mu_k, \sigma_k^2)$,

$$\mu_k = \frac{(\sum_{j \in \mathcal{D}_k} x_j)\omega_k^2 \omega_{k-1}^2 + \beta_{k-1}\omega_k^2 + \beta_{k+1}\omega_{k-1}^2}{\omega_k^2 + \omega_{k-1}^2}.$$

$$\sigma_k^2 = \frac{\omega_k^2 \omega_{k-1}^2}{\omega_k^2 + \omega_{k-1}^2},$$

para $k = 1, \ldots, J$, y $\beta_{J+1} = 0$. Para obtener el modelo de Cox, se asume que $\beta_k = \beta$ para todo k. La distribución condicional de $[t_i|(\beta, \lambda), \mathcal{D}_{obs}]$ es una exponencial a trozos truncada con parámetros $\exp(x_i \beta_k)\lambda_k$, y está dada

por

$$f(t_i|\beta, \lambda, \mathcal{D}_{obs}) = \left[1 - \exp\left\{-\sum_{l=l_i+1}^{r_i} \lambda_l \exp(x_i\beta_l)\tilde{\Delta}_l\right\}\right]^{-1} \lambda_k \exp(x_i\beta_k)$$
$$\times \exp\left\{-\sum_{l=l_i+1}^{k-1} \lambda_l \exp(x_i\beta_l)\tilde{\Delta}_l - \lambda_k \exp(x_i\beta_k)(t_i - a_{k-1})\right\}, \quad (9.6.4)$$

para $t_i \in I_k$, $l_i + 1 \leq k \leq r_i$, y $\tilde{\Delta}_l = a_l - a_{l-1}$. Note que (9.6.4) es el producto de una multinomial y una densidad exponencial truncada.

La distribución condicional de $(t_i|\beta, \lambda)$ para el modelo de Cox, cuando $\beta_k \equiv \beta$, es similar a (9.6.4).

Para el proceso de estimación e inferencia sobre el modelo de Cox generalizado se requiere adaptar y desarrollar el muestreador de Gibbs. Ibrahim, Chen y Sinha [34, p. 66] desarrollan un ejemplo sobre cáncer de mama para este modelo de Cox generalizado.

Ejemplo 9.6.1. En un estudio sobre carcinogénesis [37, pp. 1-2], se muestra el registro del tiempo hasta la muerte a partir de la inducción vaginal en ratas con el carcinógeno DMBA. Se conformaron dos grupos de acuerdo con un régimen de pretratamiento para cada uno. Cuatro ratas murieron por otras causas, por lo tanto, sus tiempos de sobrevida son censurados.

El paso DATA del paquete SAS muestra el conjunto de datos de las ratas (Rats), el cual contiene la variable *días* (el tiempo de supervivencia en días), la variable *estado* (el indicador de censura; es 0 si está censurada y 1 si no está censurada) y la variable *grupo* (el indicador del grupo de pretratamiento).

Se emplea el siguiente código de PROC PHREG, del paquete SAS, para desarrollar un análisis bayesiano del modelo exponencial a trozos. En la sintaxis del modelo (*model*) se tiene como única covariable el grupo (*Group*). En la declaración BAYES, la opción PIECEWISE estipula un modelo exponencial a segmentos o trozos, y PIECEWISE = HAZARD solicita que los riesgos constantes se modelen en la escala original. Por defecto, se utilizan ocho intervalos de riesgos constantes, y los intervalos se eligen de modo que cada uno tenga aproximadamente el mismo número de eventos.

```
data Rats;
label Days  ='Dias desde la exposici\'on a la muerte';
input Days Status Group @@;
datalines;
143 1 0   164 1 0   188 1 0   188 1 0
```

```
190 1 0   192 1 0   206 1 0   209 1 0
213 1 0   216 1 0   220 1 0   227 1 0
230 1 0   234 1 0   246 1 0   265 1 0
304 1 0   216 0 0   244 0 0   142 1 1
156 1 1   163 1 1   198 1 1   205 1 1
232 1 1   232 1 1   233 1 1   233 1 1
233 1 1   233 1 1   239 1 1   240 1 1
261 1 1   280 1 1   280 1 1   296 1 1
296 1 1   323 1 1   204 0 1   344 0 1
;
proc phreg data=Rats;
model Days*Status(0)=Group;
bayes seed=1 piecewise=hazard statistics=(summary interval)
diagnostics=(autocorr geweke ess);
run;
```

En la tabla 9.3 se muestran los intervalos (I_k), el número de ratas en riesgo (N) y las que mueren dentro de cada intervalo (evento). La última columna nombra la tasa de riesgo (λ_k) para cada intervalo.

Tabla 9.3. Modelo exponencial a trozos en ratas con cáncer

Intervalos temporales de riesgo constante				
Intervalos temporales de riesgo constante [Inferior,	Superior)	N	Evento	Parámetro de riesgo λ_k
0	176	5	5	Lambda1
176	201.5	5	5	Lambda2
201.5	218	7	5	Lambda3
218	232.5	5	5	Lambda4
232.5	233.5	4	4	Lambda5
233.5	253.5	5	4	Lambda6
253.5	288	4	4	Lambda7
288	Infty	5	4	Lambda8

La tabla 9.4 contiene la estimación para cada uno de los 8 parámetros de riesgo ($\hat{\lambda}_k$), el error y los límites de confianza para cada parámetro al 95 %. La tabla 9.5 resume, con la media y la desviación típica, las tasas de riesgo *a posteriori* y presenta los intervalos de credibilidad tipo HPD.

La figura 9.9 contiene los gráficos de la trazabilidad del muestreador de Gibbs y la gráfica de la función *a posteriori* de λ_1. Por espacio, no se muestran

Tabla 9.4. Estimadores de máxima verosimilitud

Parámetro	DF	Estimador	Error	Límites de confianza al 95 %	
Lambda1	1	0.000953	0.000443	0.000084	0.00182
Lambda2	1	0.00794	0.00371	0.000672	0.0152
Lambda3	1	0.0156	0.00734	0.00120	0.0300
Lambda4	1	0.0236	0.0115	0.00112	0.0461
Lambda5	1	0.3669	0.1959	-0.0172	0.7509
Lambda6	1	0.0276	0.0148	-0.00143	0.0566
Lambda7	1	0.0262	0.0146	-0.00233	0.0548
Lambda8	1	0.0545	0.0310	-0.00626	0.1152
Group	1	-0.6223	0.3468	-13.020	0.0573

Tabla 9.5. Resúmenes e intervalos posteriores

Parámetro	N	Media	Desviación estándar	95 % de intervalo HPD 95 %	
Lambda1	10000	0.000945	0.000444	0.000208	0.00182
Lambda2	10000	0.00782	0.00363	0.00194	0.0152
Lambda3	10000	0.0155	0.00735	0.00341	0.0301
Lambda4	10000	0.0236	0.0116	0.00478	0.0462
Lambda5	10000	0.3634	0.1965	0.0541	0.7469
Lambda6	10000	0.0278	0.0153	0.00409	0.0580
Lambda7	10000	0.0265	0.0151	0.00421	0.0569
Lambda8	10000	0.0558	0.0323	0.00637	0.1207
Group	10000	-0.6154	0.3570	-13.379	0.0652

los respectivos gráficos para los otros 7 parámetros. No obstante, al ejecutar el código SAS anterior se pueden obtener estas gráficas.

De acuerdo con las figuras 9.9 y 9.10, se observa que la trazabilidad del proceso de los Gibbs muestra una buena mezcla de estos para todos los parámetros.

En Kaeding [36], se presentan métodos bayesianos para modelos en los que la tasa de riesgo, las covariables o ambas se modelan a través de *splines*, en tiempo discreto y continuo. Funciones base B-spline, en combinación con una penalización para evitar el sobreajuste (generalmente llamados P-splines), son los principales componentes usados para modelar. Estos métodos son desarrollados desde una perspectiva bayesiana a través de algoritmos tipo MCMC.

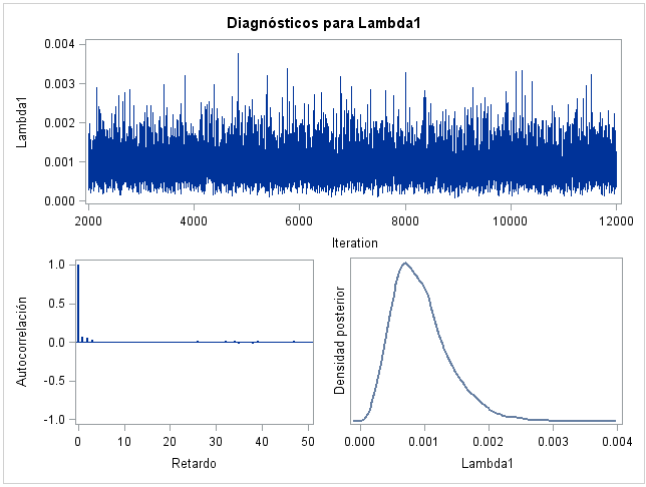

Figura 9.9. Trazabilidad del Gibbs y densidad marginal posterior de λ_1.

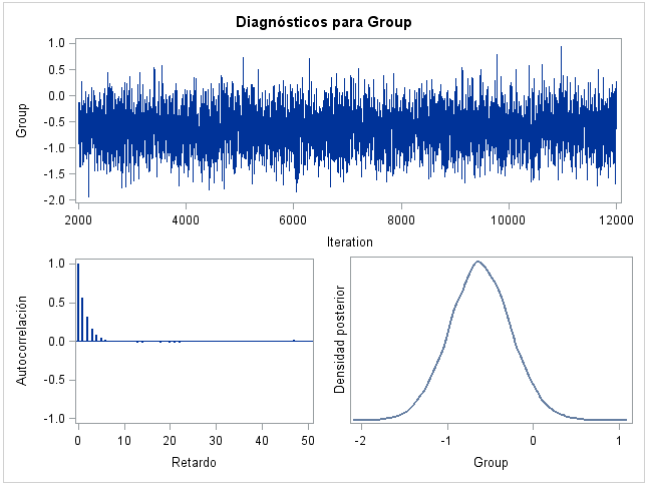

Figura 9.10. Trazabilidad del Gibbs y densidad marginal posterior de β_1.

Capítulo

diez

Modelos para datos de sobrevida discretos

$$S(t) = P(T > t) = \sum_{t_j > t} p(t_j)$$

10.1. Introducción

El tiempo en el que se produce un evento, en teoría, se ubica en cualquier punto de un intervalo de la recta numérica positiva. No obstante, el tiempo hasta un evento puede ser discreto debido a imprecisiones en la medición, a la unidad de tiempo considerada o porque el tiempo mismo es discreto. Así, por ejemplo, el evento "graduación de un estudiante" se puede presentar en: un año, un semestre académico, un mes calendario, una semana del año, un día de un mes, o un día y hora determinados. De manera que para el análisis de esta variable, su medición debería realizarse de la misma manera para todos los estudiantes incluidos en el estudio. En general, aunque se podría tener una base de datos con tiempos medidos en alguna de estas formas, esto puede no resultar práctico ni universal.

Una definición intuitiva de variable discreta es aquella en la que, entre dos de sus valores, hay un número finito de estos. De esta forma, cualquier muestra de datos es discreta. Una definición un poco más formal de variable aleatoria discreta la considera como aquella cuyos valores son un conjunto contable o enumerable.

Cualquiera de los modelos probabilísticos de tiempo continuo puede ser usado para generar un modelo discreto mediante el agrupamiento por intervalos en el eje de tiempo, como escriben Kalbfleisch y Prentice [37, pp. 46-48]. Más formalmente, el tiempo continuo está dividido en intervalos:

$$[0, t_1),\ [t_1, t_2), \ldots, [t_{q-1}, t_q), [t_q, \infty). \tag{10.1.1}$$

En resumen, los datos discretos de tiempo hasta el evento ocurren como resultados de:

- mediciones intrínsecamente discretas, donde las mediciones representan números naturales, o
- datos agrupados, que representan eventos en intervalos de tiempo subyacentes, y la respuesta se refiere a un intervalo.

En principio, los enfoques de modelamiento para datos de tiempo hasta el evento continuos también son aplicables a datos discretos de tiempo hasta el evento (al menos cuando los tiempos discretos al evento están igualmente espaciados). Sin embargo, los métodos estadísticos que han sido especialmente diseñados para tiempos discretos tienen una serie de ventajas:

- Los riesgos pueden ser formulados como probabilidades condicionales. Por lo tanto, son más accesibles para interpretación que las funciones de riesgo continuo.

- En la práctica, muchos eventos son intrínsecamente discretos o se observan en un escala discreta. En consecuencia, el uso de modelos discretos de tiempo hasta el evento es más exacto (y, por lo tanto, más apropiado) que la aproximación de los datos observados por un modelo de sobrevida continuo.
- En contraste con los modelos de sobrevida para tiempo continuo, los modelos para tiempos discretos no causan problemas con los empates.
- Los modelos de tiempo hasta el evento discretos pueden incluirse dentro del marco de los modelos lineales generalizados (MLG), como se encuentra en McCullagh y Nelder [53].

En este texto, el tiempo al evento, considerado discreto, se nota por $\mathbf{T}$, donde $\mathbf{T} = j$ significa que el evento ha ocurrido en el intervalo $[t_{j-1}, t_j)$, también llamado periodo j, con $j = 1, 2, \ldots$.

Algunos ejemplos de tiempo hasta el evento discreto son los siguientes:

1. El evento de interés es el empleo de una persona en un trabajo de tiempo parcial o un trabajo de tiempo completo. La duración del desempleo se midió en intervalos de 2 semanas. Los tiempos de eventos observados variaron de un intervalo (2 semanas) a 28 intervalos (56 semanas).
2. El número de partos de una vaca antes de que padezca la enfermedad de la mastitis.
3. El número de semestres académicos cursados por un estudiante hasta que obtenga el titulo de graduado.
4. El número de ciclos de una máquina hasta que tenga un daño.

Como se presenta en el capítulo 1, la variable aleatoria *tiempo al evento*, $\mathbf{T}$ discreta, toma los valores t_j , con $j = 1, 2, \ldots$ y con función de probabilidad $p(t_j) = P(\mathbf{T} = t_j)$, para $j = 1, 2, \ldots$, donde se asume que $t_1 < t_2 < \cdots$. La función de sobrevida para la variable aleatoria discreta $\mathbf{T}$ está dada por

$$S(t) = P(\mathbf{T} > t) = \sum_{t_j > t} p(t_j). \tag{10.1.2}$$

La función de riesgo en t_j es definida como la probabilidad condicional del evento en el tiempo t_j, dado que la unidad no ha tenido el evento al tiempo t_j. Estas se expresa como

$$h(t_j) = P(\mathbf{T} = t_j | \mathbf{T} < t_j) = \frac{p(t_j)}{S(t_j)^-}, \; i = 1, 2, \ldots, \tag{10.1.3}$$

donde $S(t_j)^- = \lim_{t \to t^-} S(t)$. La función de riesgo acumulado es

$$H(t) == \sum_{t_j < t} h(t_j). \tag{10.1.4}$$

El equivalente a la expresión (1.5.4) es

$$S(t) == \prod_{t_j \leq t} [1 - h(t_j)]. \tag{10.1.5}$$

Por las ecuaciones (10.1.3) y (10.1.5), la función de probabilidad es

$$p(t_j) = h(t_j) \prod_{j=1}^{i-1} [1 - h(t_j)]. \tag{10.1.6}$$

10.2. Modelos de regresión discretos

Las tablas de vida consideradas en el capítulo 2 producen estimaciones de tasas de riesgo discreto y funciones de sobrevida, sin asociarlas con covariables como la edad, el sexo, etc. Si hay covariables disponibles, se pueden estimar tablas de vida separadas para combinaciones de valores de covariables. Las tablas de vida así obtenidas permiten investigar el efecto de las covariables sobre la sobrevida. Aunque una falencia del método es que está restringido a casos con no demasiadas combinaciones de valores de covariables. Esta sección amplía la metodología de la tabla de vida mediante la introducción de modelos estadísticos que vinculan directamente la tasa de riesgo con una combinación aditiva de múltiples covariables.

La principal herramienta es la función de riesgo, la cual, para un vector de covariables $\mathbf{x} = (x_1; x_2, \ldots, x_p)'$, tiene la forma

$$h(t|\mathbf{x}) = P(\mathbf{T} = t|\mathbf{T} \geq t, \mathbf{x}), \ t = 1, 2, \ldots, q, \tag{10.2.1}$$

donde $h(t|\mathbf{x})$ es la probabilidad condicional de que ocurra el evento en el intervalo $[a_{t-1}; a_t)$, dado que la unidad alcanza el intervalo y se describe la tasa instantánea del evento en el momento t, ya que la unidad sobrevive hasta el tiempo t.

La correspondiente función de sobrevida es

$$S(t|\mathbf{x}) = P(\mathbf{T} > t|\mathbf{T}) = \prod_{i=1}^{t} [1 - h(i|\mathbf{x})]. \tag{10.2.2}$$

La función de sobrevida describe la probabilidad de que la falla ocurra después del momento t. En otras palabras, si se consideran los intervalos subyacentes, representa la probabilidad de sobrevida en el intervalo $[a_{t-1}; a_t)$.

Este capítulo está orientado principalmente por Tutz y Schmid [68], dado que es uno de los pocos textos dedicados al modelamiento de datos de tiempo para un evento discreto o discretizado. Un tratamiento más condensado, a partir del análisis de datos categóricos, se puede encontrar en Agresti [2] y Stokes, Davis y Koch [65].

10.2.1. Modelos de regresión paramétricos

Una forma común de especificar un modelo de datos de tiempo hasta el evento es elegir una parametrización de la función de riesgo. Para t fijo, la función de riesgo $h(t|\mathbf{x}) = P(\mathbf{T} = t|\mathbf{T} \geq t, \mathbf{x})$ modela, esencialmente, una respuesta binaria que distingue entre si el evento toma lugar en la categoría t o no (dado $\mathbf{T} \geq t$). En otras palabras, distingue entre la categoría $\{t\}$ y las categorías $\{t + 1, \ldots, k\}$ dado $\mathbf{T} \geq t$.

Mediante el uso de modelos binarios con los cuales se pueda decidir entre la categoría $\{t\}$ y las categorías $\{t + 1, \ldots, k\}$, y dado $\mathbf{T} \geq t$, se obtiene el modelo de riesgo discreto

$$h(t|\mathbf{x}) = g(\beta_{0t} + \mathbf{x}'\beta), \tag{10.2.3}$$

donde $g(\cdot)$ es una función respuesta, la cual se asume estrictamente creciente, tal como escribe en Tutz y Schmid [68, p. 37] en relación con los modelos lineales generalizados.

Se advierte que en el modelo (10.2.3) el intercepto, β_{0t}, es dependiente del tiempo. Ya que la función $g(\cdot)$ es estrictamente creciente, desde (10.2.3) se obtiene

$$g^{-1}[h(t|\mathbf{x})] = \beta_{0t} + \mathbf{x}'\beta\,. \tag{10.2.4}$$

En la terminología de los modelos lineales generalizados, $g^{-1}(\cdot)$ representa la denominada *función de enlace*.

A continuación se presentan algunos casos especiales del modelo (10.2.3).

10.2.1.1. Modelo de riesgo logístico discreto

El modelo de regresión binaria más empleado es el modelo logit, que utiliza la función de distribución logística $f(\eta) = \exp(\eta)/[1 + \exp(\eta)]$. El correspondiente modelo de riesgo logístico discreto es

$$h(t|\mathbf{x}) = \frac{\exp(\beta_{0t} + \mathbf{x}'\beta)}{1 + \exp(\beta_{0t} + \mathbf{x}'\beta)}, \tag{10.2.5}$$

que modela la ocurrencia del evento en el tiempo t (dado que se alcanza) como un modelo logístico, donde $\eta = \beta_{0t} + \mathbf{x}'\beta$ es el *predictor lineal*. Una versión alternativa es

$$\ln\left(\frac{h(t|\mathbf{x})}{1-h(t|\mathbf{x})}\right) = \beta_{0t} + \mathbf{x}'\beta\,. \tag{10.2.6}$$

Este modelo también se puede escribir como

$$\ln\left(\frac{P(\mathbf{T}=t|\mathbf{x})}{P(\mathbf{T}>t|\mathbf{x})}\right) = \beta_{0t} + \mathbf{x}'\beta\,. \tag{10.2.7}$$

La razón $P(\mathbf{T}=t|\mathbf{x})/P(\mathbf{T}>t|\mathbf{x})$ se conoce como la *razón de continuidad*, de aquí el nombre *modelo de razón de continuidad*. El modelo compara la probabilidad de ocurrencia de un evento en el tiempo t con la probabilidad de que el evento ocurra en un tiempo posterior a t.

En el modelo (10.2.6) se supone que los interceptos β_{0t} varían con el tiempo, mientras que el parámetro β es fijo. Esta separación de la variación en el tiempo y los efectos de covariables facilita la interpretación de los parámetros. Los intersectos, que varían en el tiempo, pueden ser interpretados como un riesgo de referencia o base, es decir, el riesgo que siempre está presente para cualquier conjunto dado de covariables.

Una forma alternativa del modelo dado en (10.2.7) es

$$\psi(t|\mathbf{x}) = \left(\frac{P(\mathbf{T}=t|\mathbf{x})}{P(\mathbf{T}>t|\mathbf{x})}\right) = e^{\beta_{0t}}(e^{\beta_1})_1^x \cdots (e^{\beta_p})_p^x, \tag{10.2.8}$$

donde $\psi(t|\mathbf{x})$ denota la razón de continuidad en el valor t. Se sigue que la exponencial en el intercepto es

$$e^{\beta_{0t}} = \psi(t|\mathbf{0}) = \left(\frac{P(\mathbf{T}=t|\mathbf{0})}{P(\mathbf{T}>t|\mathbf{0})}\right), \tag{10.2.9}$$

y puede interpretarse como la razón de continuidad para los valores nulos de las covariables, es decir, $\mathbf{x}' = \mathbf{0}' = (0, \ldots, 0)$. La presencia del riesgo de referencia para todos los valores de las covariables se puede ilustrar considerando el cociente de las proporciones de continuidad en dos períodos de tiempo, t y s. A partir de

$$\frac{\psi(t|\mathbf{x})}{\psi(s|\mathbf{x})} = \exp(\beta_{0t} - \beta_{0s}) \tag{10.2.10}$$

se observa que el cociente entre las razones de continuidad para dos períodos no depende de las covariables. Por lo tanto, la forma básica de las razones de continuidad es la misma para todos los valores del predictor.

La interpretación de los coeficientes β se puede considerar desde (10.2.7) y (10.2.10). El parámetro β_j representa el cambio en el logaritmo de las razones de continuidad (o el logaritmo de la razón condicional de odds) si x_j aumenta en una unidad. Así, se obtiene

$$\exp(\beta_j) = \frac{\psi(t|x_1, \ldots, x_j + 1, \ldots, x_p)}{\psi(t|x_1, \ldots, x_j, \ldots, x_p)}, \tag{10.2.11}$$

que es el factor por el cual la razón de continuidad cambia si x_j se incrementa en una unidad, es decir, cambia de x_j a x_j+1. Además, es importante advertir que el cambio no depende del tiempo.

Una visión alternativa sobre esta propiedad del modelo se obtiene al considerar dos poblaciones caracterizadas o identificadas por los valores de las covariables $\mathbf{x}$ y $\tilde{\mathbf{x}}$. Se observa, a partir de

$$\frac{\psi(t|\mathbf{x})}{\psi(t|\tilde{\mathbf{x}})} = \exp[(\mathbf{x} - \tilde{\mathbf{x}})'\beta], \tag{10.2.12}$$

que la comparación de estas subpoblaciones, en términos de la razón de continuidad, no depende del tiempo. Las subpoblaciones pueden definirse por tratamientos, métodos o grupos naturales como el sexo o la raza, por ejemplo.

10.2.1.2. Modelo de riesgo Gompertz discreto

Una distribución especial es la del mínimo valor extremo, también llamada distribución de Gompertz, $g(\eta) = 1 - \exp(\exp(\eta))$, la cual produce el modelo

$$h(t|\mathbf{x}) = 1 - \exp(-\exp(\beta_{0t} + \mathbf{x}'\beta)). \tag{10.2.13}$$

Este modelo equivale a

$$\ln(-\ln(P(\mathbf{T} > t|\mathbf{T} \geq t))) = \beta_{0t} + \mathbf{x}'\beta. \tag{10.2.14}$$

Se llama *modelo de riesgos proporcionales agrupados* porque se puede considerar como una versión discreta del modelo de riesgo proporcional de Cox. En contraste con el modelo logístico, la función de respuesta $g(\eta) = 1 \exp(-\exp(\eta))$ no es simétrica. Debido a la forma de la función de enlace, este modelo también se conoce como el complemento *log-log*.

El modelo también se puede escribir como

$$P(\mathbf{T} \leq t|\mathbf{x}) = 1 - \exp(-\exp(\tilde{\beta}_{0t} + \mathbf{x}'\beta)), \tag{10.2.15}$$

donde $\tilde{\beta}_{0t} = \ln(\sum_{i=1}^{t} \exp(\beta_{0i}))$ son las transformaciones de los parámetros de línea base.

En la tabla 10.1 se presentan varios modelos de riesgo junto con la respectiva función de enlace.

Tabla 10.1. Modelos de riesgo discretos

Modelo de riesgo / Función de enlace
Logístico $h(t\|\mathbf{x}) = \frac{\exp(\beta_{0t}+\mathbf{x}'\beta)}{1+\exp(\beta_{0t}+\mathbf{x}'\beta)}$ logit: $\ln\left(\frac{P(\mathbf{T}=t\|\mathbf{x})}{P(\mathbf{T}>t\|\mathbf{x})}\right) = \beta_{0t} + \mathbf{x}'\beta$
Probit $h(t\|\mathbf{x}) = \Phi(\beta_{0t} + \mathbf{x}'\beta)$ Probit: $\Phi^{-1}[h(t\|\mathbf{x})] = \beta_{0t} + \mathbf{x}'\beta$
Gompertz o modelo de riesgos proporcionales agrupado $h(t\|\mathbf{x}) = 1 - \exp(-\exp(\beta_{0t} + \mathbf{x}'\beta))$ clog-log: $\ln(-\ln(P(\mathbf{T} > t\|\mathbf{T} \geq t, \mathbf{x}))) = \beta_{0t} + \mathbf{x}'\beta$
Gumbel $h(t\|\mathbf{x}) = \exp(-\exp(-(\beta_{0t} + \mathbf{x}'\beta)))$ log-log: $-\ln(-\ln(h(t\|\mathbf{x}))) = \beta_{0t} + \mathbf{x}'\beta$
Exponencial $h(t\|\mathbf{x}) = \exp(\beta_{0t} + \mathbf{x}'\beta)$ log: $\ln(h(t\|\mathbf{x})) = \beta_{0t} + \mathbf{x}'\beta$

10.3. Modelo de riesgos proporcionales discreto

Ahora, se admite que el tiempo se pueda considerar en los intervalos

$$[0, a_1),\ [a_1, a_2), \ldots, [a_{q-1}, a_q), [a_q, \infty),$$

y además, se asume que se tiene el modelo de riesgos proporcionales. Si el tiempo discreto es observado, es decir, si se observa $\mathbf{T} = t$ cuando el evento ha ocurrido dentro del intervalo $[a_{t-1}, a_t)$, el *riesgo discreto*

$$h(t|\mathbf{x}) = P(\mathbf{T} = t|\mathbf{T} \geq t, \mathbf{x})$$

es de la forma

$$h(t|\mathbf{x}) = 1 - \exp(-\exp(\beta_{0t} + \mathbf{x}'\beta)), \tag{10.3.1}$$

donde los parámetros

$$\beta_{0t} = \ln(\exp(\theta_t) - \exp(\theta_{t-1})), \text{ con } \theta_t = \ln \int_0^{a_t} h_0(u)du \tag{10.3.2}$$

se obtienen a partir de la función de riesgo base $h_0(u)$.

El modelo (10.3.1) es equivalente al modelo de Gompertz o clog-log para riesgos discretos.

Se observa que el efecto de las covariables contenidas en el parámetro β es el mismo que para el modelo de Cox en tiempo continuo. Por lo tanto, si el modelo que genera los datos es el de Cox, pero los datos vienen en una forma densa y agrupados en intervalos, se puede usar el modelo clog-log descrito en la tabla 10.1 para inferir sobre el efecto de las covariables especificadas por el modelo de Cox.

Al considerar la proporcionalidad, se debe tener cuidado si se hace referencia a la proporción en tiempo continuo o en tiempo discreto, pues la proporcionalidad toma una forma diferente para la versión discreta en el tiempo. Con el modelo clog-log, la proporcionalidad no se refiere a los riesgos discretos, sino a funciones de sobrevida discretas, es decir, se obtiene

$$\ln[S(t|\mathbf{x})/S(t|\tilde{\mathbf{x}})].$$

No obstante, el modelo clog-log se denomina también como *modelo agrupado de riesgos proporcionales*. Este nombre se refiere a la conexión con el modelo de riesgos proporcionales para tiempo continuo.

10.3.0.1. Estimación

El modelo básico discreto para datos de tiempo hasta un evento

$$h(t|\mathbf{x}_i) = g(\beta_{0t} + \mathbf{x}_i'\beta)$$

es presentado en la forma

$$h(t|\mathbf{x}_i) = g(\mathbf{x}_{it}'\theta), \tag{10.3.3}$$

donde el vector de variables explicativas es $\mathbf{x}_{it}' = (0, \ldots, 0, 1, \ldots, 0, x_i')$ con 1 en el t-ésimo dígito y todos los parámetros en $\theta' = (\beta_{01}, \ldots, \beta_{0t}, \beta')$. Así, para obtener una forma cerrada simple, en $\mathbf{x}_{it}$ se incluye la dependencia del tiempo, aunque las variables $\mathbf{x}_i$ no sean dependientes del tiempo.

Considere el conjunto de datos conformado por $(t_i, \delta_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, donde n es el número de unidades, δ_i es el indicador del evento (o de la censura) sobre la unidad i, t_i es el tiempo al evento de la unidad i y $\mathbf{x}_i$, las covariables de la unidad i. Se asume que la censura y el tiempo al evento son independientes. Luego, la probabilidad de observar un tiempo al evento exacto $(t_i, \delta_i = 1)$ es dada por

$$P(\mathbf{T}_i = t_i, \delta_i = 1) = P(\mathbf{T}_i = t_i)P(t_i \leq C_i), \tag{10.3.4}$$

donde C_i es el tiempo para la censura. La probabilidad de observar censura al tiempo t_i es dada por

$$P(C_i = t_i, \delta_i = 0) = P(\mathbf{T}_i > t_i)P(C_i = t_i). \tag{10.3.5}$$

Un evento en el intervalo $[a_{t-1}, a_{t_i})$ implica que $t_i \leq C_i$, y la censura en el intervalo $[a_{t-1}, a_{t_i})$ implica que el evento ocurre después de a_{t_i}, es decir que hay sobrevida más allá de a_{t_i}.

Empleando la función de riesgo discreta condicionada a covariables y con la expresión para la función de sobrevida discreta sujeta a las mismas covariables, a partir de (10.2.2) se obtiene la contribución de la i-ésima unidad a la verosimilitud:

$$L_i \propto h(t_i|\mathbf{x}_i)^{\delta_i}[1 - h(t_i|\mathbf{x}_i)]^{1-\delta_i} \prod_{j=1}^{t_i-1}(1 - h(j|\mathbf{x}_i)). \tag{10.3.6}$$

Esta verosimilitud parece ser muy específica para un modelo discreto de tiempo hasta el evento y con censura a derecha.

Con una inspección más detallada, se observa la semejanza entre esta verosimilitud y la verosimilitud asociada con una sucesión de respuestas binarias. Por definición, para una observación no censurada ($\delta_i = 1$), la sucesión es $(y_{i1}, \ldots, y_{it_i}) = (0, \ldots, 0, 1)$. Y para una observación censurada ($\delta_i = 0$), la sucesión es $(y_{i1}, \ldots, y_{it_i}) = (0, \ldots, 0)$, por lo que la verosimilitud (10.3.6) puede escribirse como

$$L_i \propto \prod_{s=1}^{t_i} h(s|\mathbf{x}_i)^{y_{is}}(1 - h(s|\mathbf{x}_i))^{1-y_{is}}. \tag{10.3.7}$$

Por lo tanto, L_i es equivalente a la verosimilitud de un modelo de respuesta binaria con valores $y_{is} \in \{0, 1\}$. Generalmente, sobre cada unidad i, se tienen t_i de estas observaciones. Tenga en cuenta, sin embargo, que las observaciones binarias no son valores artificialmente construidos. De hecho, los valores y_{is} codifican las decisiones binarias para la transición al siguiente período. Específicamente, y_{is} codifica la transición desde el intervalo

$[a_{s-1}, a_s)$ al intervalo $[a_s, a_{s+1})$ en la forma

$$y_{is} = \begin{cases} 1, & \text{si la unidad } i \text{ tiene el evento en } [a_{s-1}, a_s) \\ 0, & \text{si la unidad } i \text{ no tiene el evento en } [a_{s-1}, a_s), \end{cases} \tag{10.3.8}$$

con $s = 1, \ldots, t_i$. Para tiempos al evento exactos, se tiene el vector de observación $(0, \ldots, 0, 1)'$ de tamaño $t_i \times 1$, es decir, no le ocurre el evento durante los primeros t_i intervalos y le ocurre el evento en el t_i-ésimo intervalo. Para observaciones censuradas, se sabe que en los primeros t_i intervalos no ha ocurrido el evento. Como consecuencia, el *logaritmo de la verosimilitud total del modelo de supervivencia discreta* está dado por

$$l \propto \sum_{i=1}^{n} \sum_{s=1}^{t_i} y_{is} \ln h(s|\mathbf{x}_i) + (1 - y_{is}) \ln(1 - h(s|\mathbf{x}_i)). \tag{10.3.9}$$

Para los cálculos, se deben generar observaciones binarias antes de ajustar el modelo binario para transiciones.

A manera de ilustración, siguiendo a Tutz y Schmid [68, pp. 54-57], considere el modelo de riesgo discreto con predictor lineal $\eta_{ir} = \beta_{0r} + \mathbf{x}_i'\beta$. Una forma condensada del predictor lineal es $\mathbf{x}_{ir}'\theta$, con vector de parámetros $\theta = (\beta_{01}, \ldots, \beta_{0q}, \beta')$. Para predictores lineales de respuesta binaria se deben adaptar las covariables a las observaciones bajo consideración. Para la generación de la matriz de diseño, se debe distinguir entre observaciones completas y observaciones censuradas.

Si $\mathbf{T} = t_i$ y $\delta_i = 1$, las observaciones binarias t_i y las variables de diseño son dadas por la matriz de datos aumentada:

Obs. binarias	Variables de diseño					
0	1	0	0	$\cdots$	0	$\mathbf{x}_i'$
0	0	1	0	$\cdots$	0	$\mathbf{x}_i'$
0	0	0	1	$\cdots$	0	$\mathbf{x}_i'$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
1	0	0	0	$\cdots$	1	$\mathbf{x}_i'$

De forma similar, si $\mathbf{T} = t_i$ y $\delta_i = 0$, las observaciones binarias t_i y las variables de diseño son dadas por la matriz de datos aumentada:

Obs. binarias	Variables de diseño					
0	1	0	0	$\cdots$	0	$\mathbf{x}_i'$
0	0	1	0	$\cdots$	0	$\mathbf{x}_i'$
0	0	0	1	$\cdots$	0	$\mathbf{x}_i'$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
0	0	0	0	$\cdots$	1	$\mathbf{x}_i'$

En lugar de construir la matriz de diseño, también se pueden construir las observaciones con tiempo transcurrido t, el cual debe considerarse como un factor cuando se utiliza alguna herramienta computacional. Así, para $\mathbf{T} = t_i$ y $\delta_i = 1$ se emplea la siguiente codificación:

Obs. binarias	Variables de diseño	
0	1	$\mathbf{x}_i'$
0	2	$\mathbf{x}_i'$
0	3	$\mathbf{x}_i'$
$\vdots$	$\vdots$	$\vdots$
1	t_i	$\mathbf{x}_i'$

De manera similar, para $\mathbf{T} = t_i$ y $\delta_i = 0$ se emplea la siguiente codificación:

Obs. binarias	Variables de diseño	
0	1	$\mathbf{x}_i'$
0	2	$\mathbf{x}_i'$
0	3	$\mathbf{x}_i'$
$\vdots$	$\vdots$	$\vdots$
0	t_i	$\mathbf{x}_i'$

10.3.0.2. Modelamiento del tiempo al evento con censura

Hasta ahora la contribución de la censura a la verosimilitud ha sido ignorada. No obstante, el proceso de censura también puede ser considerado e incluido en el modelo de riesgo discreto. Para este propósito, se nota con

$$h_{cen}(t|\mathbf{x}_i) = P(C = t|C \geq t, \mathbf{x}_i)$$

la función de riesgo para el tiempo de censura. Con la censura al final del intervalo, la contribución de la i-ésima unidad a la verosimilitud, como se consideró antes en (10.3.6), es

$$L_i = c_i h(t_i|\mathbf{x}_i)^{\delta_i}[1 - h(t_i|\mathbf{x}_i)]^{1-\delta_i} \prod_{j=1}^{t_i-1}(1 - h(j|\mathbf{x}_i)),$$

donde $c_i = P(C_i \geq t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i}$ representa la contribución de las observaciones censuradas. Una expresión equivalente para c_i es

$$c_i = h_{cen}(t|\mathbf{x}_i)^{1-\delta_i} \prod_{s=1}^{t_i-1}(1 - h_{cen}(s|\mathbf{x}_i)).$$

Si $\delta_i = 0$, se obtiene

$$c_i = \prod_{s=1}^{t_i-\delta_i} h_{cen}(s|\mathbf{x}_i)^{y_{is}^{(c)}} (1 - h_{cen}(s|\mathbf{x}_i))^{1-y_{is}^{(c)}}.$$

Con respecto a la matriz de diseño, se obtiene, para el proceso de censura desarrollando el tiempo t $\delta_i = 0$, las t_i observaciones:

Obs. binarias	Variables de diseño	
0	1	$\mathbf{x}_i'$
0	2	$\mathbf{x}_i'$
0	3	$\mathbf{x}_i'$
⋮	⋮	⋮
1	t_i	$\mathbf{x}_i'$

Para el proceso de censura con tiempo de ejecución t y $\delta_i = 0$, se obtienen las $t_i - 1$ observaciones:

Obs. binarias	Variables de diseño	
0	1	$\mathbf{x}_i'$
0	2	$\mathbf{x}_i'$
0	3	$\mathbf{x}_i'$
⋮	⋮	⋮
0	$t_i - 1$	$\mathbf{x}_i'$

Para la estimación mediante la máxima verosimilitud, que involucre tiempo al evento y el proceso de censura, se usa el logaritmo de la verosimilitud $l = l_T + l_C$, donde

$$l_T = \sum_{i=1}^{n} \sum_{s=1}^{t_i} (y_{is} \ln h(s|\mathbf{x}_i) + (1 - y_{is}) \ln(1 - h(s|\mathbf{x}_i)))$$

es el logaritmo de la verosimilitud que depende del riesgo discreto al evento (es decir sobre $h(s|\mathbf{x}_i)$), y

$$l_C = \sum_{i=1}^{n} \sum_{s=1}^{t_i - \delta_i} (y_{is}^{(c)} \ln h_{cen}(s|\mathbf{x}_i) + (1 - y_{is^{(c)}}) \ln(1 - h_{cen}(s|\mathbf{x}_i)))$$

es el logaritmo de la verosimilitud que depende del riesgo discreto a la censura (es decir, sobre $h_{cen}(s|\mathbf{x}_i)$). Si el tiempo para el evento y la censura no comparten parámetros, l_T y l_C se pueden maximizar por separado.

10.3.0.3. Censura al comienzo del intervalo

La derivación del log-verosimilitud en las secciones anteriores se basa en el supuesto de que la censura ocurre al final del intervalo. Las definiciones y las derivaciones cambian si se asume la censura al comienzo del intervalo. En este caso se tiene

$$\delta_i = \begin{cases} 1, & \text{si } \mathbf{T}_i < C_i \\ 0, & \text{si } \mathbf{T}_i \geq C_i \end{cases}$$

y

$$P(\mathbf{T}_i = t_i, \delta_i = 1) = P(\mathbf{T}_i = t_i) P(C_i > t_i)$$
$$P(\mathbf{T}_i = t_i, \delta_i = 0) = P(\mathbf{T}_i \geq t_i) P(C_i = t_i).$$

La contribución de la i-ésima unidad a la verosimilitud está dada por

$$L_i = P(\mathbf{T}_i = t_i)^{\delta_i} P(\mathbf{T}_i \geq t_i)^{1-\delta_i} P(C_i > t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i}. \quad (10.3.10)$$

Sin censura, la verosimilitud es igual a

$$L_i = c_i h(t_i|\mathbf{x}_i)^{\delta_i} \prod_{j=1}^{t_i - 1} (1 - h(j|\mathbf{x}_i)). \quad (10.3.11)$$

Por definición, la sucesión $(y_{i1}, \ldots, y_{it_i}) = (0, \ldots, 0, 1)$ para una observación no censurada ($\delta_i = 1$). La verosimilitud puede escribirse como

$$L_i \propto \prod_{s=1}^{t_i} h(s|\mathbf{x}_i)^{y_{is}} (1 - h(s|\mathbf{x}_i))^{1-y_{is}}. \quad (10.3.12)$$

Para una observación censurada se define $(y_{i1}, \ldots, y_{it_i-1}) = (0, \ldots, 0)$, así, se obtiene

$$L_i \propto \prod_{s=1}^{t_i-1} h(s|\mathbf{x}_i)^{y_{is}}(1-h(s|\mathbf{x}_i))^{1-y_{is}}. \tag{10.3.13}$$

De forma cerrada, la contribución de la i-ésima unidad a la verosimilitud es

$$L_i \propto \prod_{s=1}^{t_i-(1-\delta_i)} h(s|\mathbf{x}_i)^{y_{is}}(1-h(s|\mathbf{x}_i))^{1-y_{is}}, \tag{10.3.14}$$

y el logaritmo de la verosimilitud completa para el modelo es

$$l \propto \sum_{i=1}^{n} \sum_{s=1}^{t_i-(1-\delta_i)} (y_{is} \ln h(s|\mathbf{x}_i) + (1-y_{is}) \ln(1-h(s|\mathbf{x}_i))).$$

El logaritmo de la verosimilitud es nuevamente equivalente al logaritmo de la verosimilitud de un modelo para respuesta binaria, el cual se escribe como $P(y_{ij} = 1|\mathbf{x}_i) = g(\mathbf{x}_i\beta)$ con observaciones binarias

$$y_{11}, \ldots, y_{1,t_1-(1-\delta_1)}, y_{21}, \ldots, y_{2,t_2-(1-\delta_2)}, \ldots, y_{n1}, \ldots, y_{n,t_n-(1-\delta_n)}.$$

En total, se tienen $\sum_i (t_i - 1 + \delta_i)$ observaciones binarias, donde la suma varía con las observaciones censuradas y también depende de los tiempos al evento completos.

10.3.0.4. Errores estándar

Consideremos el caso en el que la censura ocurre al final del intervalo. Entonces, la log-likelihood puede usarse para derivar errores estándar aproximados para las estimaciones de coeficientes de un modelo de riesgo discreto. Mediante (10.3.15) se deriva la matriz de información, y esta es dada por

$$\mathcal{I}(\beta) = E\left(-\frac{\partial^2 l(\beta)}{\partial\beta\,\partial\beta'}\right) = \sum_{i=1}^{n}\sum_{s=1}^{t_i} \mathbf{x}_{it}\mathbf{x}_{it}'\left(\frac{\partial g(\eta_{it})}{\partial\eta}\right)^2 / \sigma_{it}^2, \tag{10.3.15}$$

donde $\eta_{it} = \mathbf{x}_{it}\beta$ y $\sigma_{it}^2 = g(\eta_{it})(1-g(\eta_{it}))$. Una aproximación a los errores estándar se obtiene a partir de la distribución asintótica del estimado de máxima verosimilitud $\hat{\beta}$, esta es

$$\hat{\beta} \xrightarrow{D} N(\beta, \mathcal{I}(\beta)^{-1}). \tag{10.3.16}$$

Con esta convergencia en distribución de $\hat{\beta}$, se obtienen los errores estándar de cada uno de sus componentes, y en consecuencia, se pueden construir intervalos de confianza de Wald en el nivel $(1-\alpha)$ para cada β_k. Este intervalo es

$$\hat{\beta}_k \mp Z_{1-\alpha/2}\sqrt{\mathcal{I}(\beta)^{-1}_{kk}}). \tag{10.3.17}$$

Capítulo
once

Tamaño de muestra en análisis de sobrevida

$$Var(\hat{\delta}) = \frac{1}{np(1-p)}(p\pi_0 + (1-p)\pi_1)$$

11.1. Introducción

En estudios de análisis de tiempo hasta la ocurrencia de un evento o de sobrevida es importante conocer, en lo posible antes de iniciar el estudio, *el número de unidades* (personas, animales u objetos) y *el número de eventos* requeridos para que el estudio tenga unos niveles de validez y confiabilidad en términos probabilísticos. El número de unidades y el número de eventos (a lo sumo, uno sobre cada unidad) son cantidades correlacionadas.

Decidir cuántas unidades se deben incluir en un estudio sobre tiempo hasta la ocurrencia de un evento es un elemento importante, por ejemplo, en el diseño de un experimento clínico, industrial, pedagógico o social. En el marco clásico de verificación de hipótesis, para cualquier tipo de resultado se debe especificar el cambio de efecto que se busca (diferencia), la variabilidad en la estadística de verificación, el nivel de significación de la prueba, la potencia deseada de la prueba para detectar un mínimo cambio del efecto (la diferencia significativa) e, incluso, la disponibilidad de recursos económicos y logísticos para desarrollar el estudio.

En el análisis de datos de tiempo hasta el evento hay factores adicionales que se deben especificar con respecto al mecanismo de censura y la distribución de sobrevida particular, bajo las hipótesis nula y alterna, respectivamente. Primero, se requiere especificar qué modelo paramétrico de la variable *tiempo hasta el evento* se está asumiendo, o qué prueba semiparamétrica se va a emplear, por ejemplo, la prueba tipo log-rank. En la estimación del tamaño de muestra para la comparación de medias, se asume la distribución normal. De manera análoga, la estimación del tamaño de muestra (el número de eventos o el número de unidades) en análisis de sobrevida se hace sobre el supuesto de la variable *tiempo al evento* distribuida exponencialmente. Esto permite determinar tanto el número de eventos como el número de unidades requerido para cumplir con la potencia y con otras especificaciones del diseño. En segundo lugar, se debe proporcionar, por razones administrativas y logísticas, una estimación de la cantidad de unidades que se necesitan incluir en el estudio para producir el número requerido de eventos. A continuación se presenta lo relacionado con el tamaño de la muestra en datos de tiempo para evento; esto apoyado en Moore [55, cap. 11] y en Collet [10, cap. 10].

11.2. Potencia y tamaño de muestra en estudios a una vía de clasificación

Para estudios con un único grupo, se asume que la variable *tiempo hasta el evento* se ajusta a un modelo exponencial, con función de riesgo $h(t, \theta) = \theta$ y distribución de sobrevida $S(t, \theta) = e^{-\theta t}$. Se quiere verificar la hipótesis $H_0 : \theta = \theta_0$ frente a $H_a : \theta = \theta_a$. Bajo la hipótesis nula, la media es $1/\theta_0$, y bajo el riesgo alternativo, la media es $1/\theta_a$. Estas pueden corresponder al riesgo para un grupo (por ejemplo, *tratamiento*) y el riesgo para otro grupo (por ejemplo, *control*). Así, la razón de medias, que equivale a la razón de riesgos, se escribe como $\Delta = \mu_a/\mu_0 = \theta_0/\theta_a$. Una vez que se completa el estudio, se obtienen los tiempos hasta el evento independientes $t_1, \ldots, t_n$ junto con los respectivos indicadores de censura $\delta_1, \ldots, \delta_n$, donde n es el número de unidades incluidas en el estudio o ensayo.

El objetivo final es determinar cuántas unidades se requieren para detectar una cierta razón de riesgo, con una potencia y un nivel de significancia establecidos. El proceso de derivación de la fórmula es similar al seguido en pruebas sobre la media de observaciones con distribución normal, aunque algunos ajustes son necesarios para dar cuenta de la presencia de censura. El ajuste principal es que la fórmula de tamaño de muestra especificará directamente **e**, el número de eventos necesarios para lograr la potencia preestablecida. Una vez que se tiene **e**, de manera separada se usa un método para encontrar el número de unidades, **n**, necesarias para producir el **e** requerido.

En la práctica, las hipótesis nula y la alterna pueden expresarse en términos de las respectivas medianas de sobrevida o de la probabilidad de sobrevida en un tiempo específico t. Una mediana del tiempo de sobrevida puede transformarse en una tasa de riesgo al expresar $1/2 = e^{-\theta t}$ como $\theta = \ln(2)/t$. De manera similar, una tasa de supervivencia p en el tiempo t puede escribirse como $p = e^{-\theta t}$, que también se puede expresar como $\theta = -\ln(p)/t$. Estas comparaciones son estrictamente válidas solo si la distribución de sobrevida es exponencial. Sin embargo, para otras distribuciones de sobrevida, las conversiones entre la mediana de sobrevida y las tasas de supervivencia pueden proporcionar aproximaciones cuando las tasas de sobrevida están cerca del 50 %.

La forma más directa de derivar una fórmula para el tamaño de muestra se basa en una estadística de Wald, pero la fórmula resultante varía según la parametrización de la verosimilitud. La derivación más simple usa el parámetro $\lambda = \ln(\mu) = -\ln(\theta)$. Entonces, se puede expresar la función

logaritmo de la verosimilitud de la siguiente manera:

$$l(\lambda) = e\ln(\theta) - \theta E = -\lambda e - E\exp^{-\lambda}. \tag{11.2.1}$$

De la estimación máxima verosímil se sigue que $\hat{\lambda} = \ln(E)/e$ y su varianza es $var(\hat{\lambda}) \approx 1/e$, donde $e = \sum_{i=1}^{n} \delta_i$ y $E = \sum_{i=1}^{n} t_i$ son el número total de eventos y el total de tiempo de las unidades, respectivamente.

Sea Z_α el punto crítico sobre una distribución normal estándar que determina la región crítica a la derecha de tamaño α. Así

$$Z_\alpha = \frac{k - \lambda_0}{1/\sqrt{e}},$$

por lo tanto,

$$k = \lambda_0 + \frac{Z_\alpha}{\sqrt{e}}. \tag{11.2.2}$$

Ahora, bajo la hipótesis alterna, $H_a = \lambda_a$, la potencia de la estadística está dada por

$$1 - \beta = P\left(\hat{\lambda} > k | \lambda = \lambda_a\right) = P\left(Z > \frac{k - \lambda_a}{1/\sqrt{e}}\right),$$

que equivale a

$$Z_{1-\beta} = -Z_\beta = \sqrt{e}(k - \lambda_a).$$

Reemplazando el valor de k, a partir de (11.2.2), se obtiene

$$-Z_\alpha = \sqrt{e}\left(\lambda_0 + \frac{Z_\alpha}{\sqrt{e}} - \lambda_a\right).$$

Resolviendo para e, de acuerdo con la definición de Δ, resulta

$$e = \frac{(Z_\alpha + Z_\beta)^2}{(\lambda_a - \lambda_0)^2} = \frac{(Z_\alpha + Z_\beta)^2}{(\ln \Delta)^2}. \tag{11.2.3}$$

La expresión (11.2.3) suministra el número de eventos mínimo, **e**, necesarios para alcanzar la potencia y significancia especificada. No es la cantidad de unidades expuestas al evento.

Un aspecto de esta prueba es que diferentes parametrizaciones conducen a fórmulas algo diferentes para el tamaño de muestra (**e**). Esta dependencia de la parametrización puede ser obviada si se trabaja directamente con la estadística de razón de verosimilitud.

Para **e** fijo, $E = \sum_{i=1}^{n} t_i$ tiene una distribución gama con índice **e** y parámetro de escala θ. Cox [14, p. 35] demuestra que

$$\mathcal{W} = \frac{2e\alpha}{\hat{\alpha}} \sim \chi^2_{se},$$

este resultado es aproximadamente válido para cualquier esquema de censura. Bajo $H_0 : \theta = \theta_0$, con un nivel de significancia α y un máximo error Tipo II igual a β (potencia igual a $1 - \beta$), se tiene

$$\Delta = \frac{\theta_0}{\theta_a} = \frac{\chi^2_{2e,\alpha^*}}{\chi^2_{2e,1\beta}}. \tag{11.2.4}$$

Para valores específicos de significancia α, potencia $1-\beta$ y razón Δ, se puede resolver (11.2.4) para encontrar el número de eventos e.

Moore [55, p. 160] desarrolla el siguiente código R, de acuerdo con la expresión (11.2.3), para calcular el número de eventos e requeridos con una significancia, potencia y razón especificados.

```
expLogMeanDeaths <- function(Delta, alpha, pwr) {
z.alpha <- qnorm(alpha, lower.tail=F)
z.beta <- qnorm(1-pwr, lower.tail=F)
num <- (z.alpha + z.beta)^2
denom <- (log(Delta))^2
ee <- num/denom
ee }
```

De manera similar, Moore [55, pp. 160-161] presenta el código R para el cálculo de **e** alternativo contenido en (11.2.4). Así, primero se calcula Δ mediante

```
expLikeRatio <- function(e, alpha, pwr) {
num <- qchisq(alpha, df=(2*e), lower.tail=F)
denom <- qchisq(pwr, df=(2*e), lower.tail=F)
Delta <- num/denom
Delta }
```

Para obtener el número de eventos **e** para un Δ especificado, se define internamente la función "LRD", la cual es cero en el número requerido de eventos:

```
expLRdeaths <- function(Delta, alpha, pwr) {
LRD <- function(x, alpha, pwr)
expLikeRatio(x, alpha, pwr) - Delta
result <- uniroot(f=LRD, lower=1,upper=1000,
alpha=alpha, pwr=pwr)
result$root }
```

Suponga que se está diseñando un ensayo de oncología de fase II en el que se planifica una prueba con un nivel de significancia del 5 % (prueba unilateral) y con una potencia del 80 % para detectar una razón de riesgo de $\Delta = 1.5$. Una vez las funciones anteriores se han ingresado en R, podemos encontrar el número requerido de eventos de la siguiente manera:

```
> expLRdeaths(1.5, 0.05, 0.8)
[1] 36.33916
> expLogMeanDeaths(1.5, 0.05, 0.8)
[1] 37.60635
```

Es decir, se requiere, redondeando, de 38 eventos de acuerdo con el método de log mean (función "expLogMeanDeaths") y 37 eventos según el método de razón de verosimilitud (función "expLikeRatio"). Los resultados de los dos métodos son similares.

11.3. Determinación de la probabilidad del evento

En muchos estudios, usualmente se requiere conocer el número de unidades a incluir y no el número de eventos que ocurren. Así, se necesita tener una estimación de la proporción, π, de unidades a las que les ocurrirá el evento durante el tiempo del estudio. Si todas las unidades ingresaron, se tendría $\pi = S(t, \theta)$, donde t es el tiempo de seguimiento. Sin embargo, las unidades realmente ingresan durante un período de acumulación de longitud a (acumulación de unidades), y luego, después de la acumulación de unidades en el estudio o ensayo, las unidades son seguidas durante un tiempo adicional f. Entonces, una unidad que ingrese en el momento $t = 0$ tendrá una probabilidad de que el evento ocurra $\pi(0) = 1 - S(a + f, \theta)$, es decir que una unidad que ingresa en el momento $t = 0$ tendrá el máximo tiempo posible de seguimiento igual a $a + f$. Pero una unidad que entra al estudio en el momento a, que es el final del período de acumulación de unidades (reclutamiento), tendrá un tiempo mínimo posible de seguimiento f. Por lo tanto, esa unidad tendrá una probabilidad de que ocurra el evento igual a $\pi(a) = 1 - S(f, \theta)$.

Si las unidades ingresaron en números iguales en los momentos 0 o a solamente, entonces la probabilidad del evento sería $(\pi(0) + \pi(a))/2$.

Un escenario más realista es aquel en el que las unidades ingresan de manera uniforme entre los tiempos 0 y a, de modo que el ingreso de las unidades tiene una distribución uniforme $U(0, a)$. Así, la probabilidad del

evento, π, se obtiene con el promedio de estos dos tiempos. Por esto, una unidad que ingresa en un tiempo t es seguida durante un tiempo adicional $a+f-t$. Esta idea se expresa en la siguiente integral de seguimiento, la cual utiliza el hecho de que la probabilidad de ocurrencia del evento, dado que la unidad entra en el tiempo t, es $1-S(a+f-t,\theta)$:

$$\begin{aligned}\pi &= \int_0^a \frac{1}{a} P(Evento|\text{ingresa en el tiempo } t)dt \\ &= \int_0^a \frac{1}{a}(1-S(a+f-t,\theta))dt.\end{aligned}$$

Empleando la transformación $u = a+f-t$ y considerando que para una distribución exponencial $S(u,\theta) = e^{-\theta u}$, por integración se tiene que

$$\begin{aligned}\pi &= 1-\frac{1}{a}\int_f^{a+f} S(u,\theta)du \\ &= 1-\frac{1}{a}\int_f^{a+f} e^{-\theta u}du \\ &= 1-\frac{1}{a\theta}\left\{e^{-\theta f}-e^{-\theta(a+f)}\right\}.\end{aligned} \tag{11.3.1}$$

Mediante la siguiente función en código R se emplea la última expresión para calcular la probabilidad del evento:

```
prob.death <- function(theta, accrual, followup) {
probDeath <- 1 - (1/(accrual*theta))*
(exp(-theta*followup) - exp(-theta*(accrual + followup)))
probDeath
}
```

Para el ejemplo anterior, se planea una sola muestra de un ensayo clínico con una prueba a un nivel de significancia del 5 % (unilateral), y se requiere un 80 % de potencia para detectar una razón de riesgos de 1.5. Suponga que la tasa de hipótesis nula es $\theta_0 = 0.15$, de modo que la tasa de riesgo de la hipótesis alternativa es $\theta_1 = \theta_0/\Delta = 0.15/1.5 = 0.10$. Suponga ahora que el período de acumulación es $a = 2$ años y que el período de seguimiento adicional es $f = 3$ años. En el ejemplo anterior, se encontró que se requería 38 eventos (muertes). Para obtener una estimación del número de pacientes necesarios para producir este número de eventos (muertes), primero se calcula la probabilidad del evento (muerte) bajo $H_1 : \theta = 0.10$.

```
> prob.death(lambda=0.10, accrual=2, followup=3)
[1] 0.3285622
```

Entonces, el número de pacientes necesario para el ensayo clínico es aproximadamente igual a $38/0.3285622 = 115.6 \approx 116$. Esta estimación depende no solo del supuesto de una distribución exponencial, sino también del valor desconocido del parámetro θ de la distribución exponencial. Usar un valor específico, como $\theta_1 = 0.10$, es útil en la etapa de planificación del ensayo, ya que se necesita una estimación del número de pacientes.

11.4. Tamaño de muestra para comparar dos distribuciones de sobrevida

Se considera un estudio (ensayo o experimento) comparativo donde un tratamiento experimental se compara con un tratamiento control o estándar. Se quiere verificar la hipótesis nula $H_0 : \ S_0 \geq S_1$ frente a la alternativa $H_a : \ S_0 < S_1$ para todo t, donde S_0 y S_1 son distribuciones de sobrevida exponencial con riesgos θ_0 y θ_1 para el grupo control y el grupo tratado, respectivamente. Para determinar el tamaño de muestra requerido, se considera la razón de riesgo $\Delta = \frac{\theta_0}{\theta_1}$. Se supone que el último riesgo (θ_1) es el más pequeño y que la prueba puede considerarse unilateral de la forma $H_0 : \ \Delta = 1$ frente a $H_a : \ \Delta > 1$.

Suponga que p representa la proporción de unidades aleatoriamente asignadas al grupo de control. Por lo general, se usa la asignación proporcionalmente igual, de manera que $p = 0.5$, pero esto no es necesario. Se denota con n el número total de unidades en el estudio, y se tiene que $n_0 = np$ y $n_1 = n(1-p)$ es el número de unidades en el grupo control y en el grupo experimental, respectivamente. Cuando se haya completado la prueba, se observarán $\mathbf{e}_0$ y $\mathbf{e}_1$ eventos en el grupo control y experimental, respectivamente. Y el total del tiempo al evento será $V_0 = \sum_i t_{0i}$ y $V_1 = \sum_i t_{1i}$ para los grupos control y experimental, respectivamente.

Los estimadores de máxima verosimilitud de los riesgos son $\theta_0 = e_0/V_0$ y $\theta_1 = e_1/V_1$. Para comparar las dos distribuciones es más conveniente usar $\delta = \ln \Delta = \ln \theta_0 - \ln \theta_1$.

Sean π_0 y π_1 las probabilidades al evento en el grupo control y tratamiento, respectivamente. Así,

$$var(\hat{\delta}) = \frac{1}{np(1-p)} \frac{p\pi_0 + (1-p)\pi_1}{\pi_0 \pi_1}.$$

Si se define,

$$\tilde{\pi} = \frac{p\pi_0 + (1-p)\pi_1}{\pi_0 \pi_1} = \left(\frac{p}{\pi_1} + \frac{1-p}{\pi 0} \right)^{-1}.$$

La prueba de la hipótesis se asume con un nivel de significancia α y una potencia $1-\beta$. Con lo anterior y a partir de la ecuación (11.2.3), Moore [55, p.164] encuentra el tamaño de muestra

$$n = \frac{(Z_\alpha + Z_\beta)^2}{p(1-p)\tilde{\pi}} = \frac{e}{\tilde{\pi}}. \tag{11.4.1}$$

11.5. Tamaño de muestra para comparar dos distribuciones mediante log-rank

Cuando se analizan los datos de estudios comparativos (grupos G_1 y G_2), usualmente se emplea una estadística de verificación tipo log-rank. Por lo tanto, se desarrolla una expresión para el tamaño de muestra basada en la estadística log-rank. En esta estadística, que se fundamenta en el supuesto de riesgos proporcionales en los dos grupos, la razón $\Delta = \theta_1(t)/\theta_2(t)$ es constante. Por esto se emplea la estadística log-rank, LR, y su varianza, como se muestra en (2.7.1), que al i-ésimo tiempo al evento está dada por

$$\sigma^2_{e_{j2}} = \frac{n_{j1}n_{j2}e_j(n_j - e_j)}{n_j^2(n_j - 1)},$$

con n_j siendo el número de unidades expuestas al evento (en riesgo) en un tiempo inmediatamente inferior a $t_{(j)}$.

Tabla 11.1. Núm. de eventos por grupo al momento $t_{(j)}$

	Grupos		
Evento (E)	1	2	Total
Ocurre	e_{j1}	e_{j2}	e_j
No ocurre	$n_{j1} - e_{j1}$	$n_{j2} - e_{j2}$	$n_j - e_j$
Total	n_{j1}	n_{j2}	n_j

Se considera una tabla 2×2, como se muestra en la tabla 11.1, tomando como referencia el grupo 2 (igual con el grupo 1). Se supone que el número de eventos ocurridos en cada tiempo al evento es pequeño en relación con el número de unidades en riesgo, y que la proporción $p \approx n_{2j}/n_j$ de unidades asignadas al grupo 2 es constante en el tiempo. Bajo estas condiciones, se

tiene la siguiente aproximación:

$$\sigma^2_{e_{j2}} = \frac{n_{j1}n_{j2}e_j(n_j - e_j)}{n_j^2(n_j - 1)} \approx \frac{n_{j1}n_{j2}e_j}{n_j^2} \approx p(1-p)e_j.$$

La varianza de la estadística log-rank es aproximadamente

$$V_2 = \sum_{j=1}^{E} \sigma^2_{e_{j2}} \approx p(1-p)\sum_{j=1}^{E} e_j \approx p(1-p)e.$$

Para encontrar el número de eventos necesarios que detecten una diferencia entre los grupos (debido a los tratamientos o a otro factor), la "diferencia" a detectar es $\delta = \ln(\Delta) = \ln[\theta_1(t)/\theta_2(t)]$, con una potencia $1-\beta$ y con una significancia α para una estadística de verificación a dos colas tipo log-rank. Este número es dado por la siguiente fórmula:

$$e = \frac{(Z_{\alpha/2} + Z_\beta)^2}{p(1-p)\delta^2}. \tag{11.5.1}$$

Si se quiere el mismo número de pacientes en ambos grupos, entonces

$$e = \frac{(Z_{\alpha/2} + Z_\beta)^2}{\delta^2/4}. \tag{11.5.2}$$

Esta expresión suministra el número de eventos, no el número de unidades. Para encontrar el número de unidades se requiere estimar la probabilidad del evento, de lo cual se obtiene la expresión (11.4.1) para el cálculo del número de unidades.

11.6. Probabilidad del evento desde una estimación no paramétrica de la función de sobrevida

Una forma de hallar el número de unidades necesarias para producir un número determinado de eventos es asumir que las unidades ingresan uniformemente sobre el periodo de acumulación y que la sobrevida es derivada de una distribución exponencial, como se trata en la sección 11.3. Si la distribución de sobrevida es estimada para uno de los grupos, sin pérdida de generalidad, suponga que es estimada para el grupo control, entonces, se puede asumir que los dos grupos tienen un riesgo proporcional en el tiempo. Así se estima una probabilidad del evento en cualquiera de los grupos

(por la proporcionalidad del riesgo) sin asumir que la distribución es exponencial.

La función de sobrevida para el grupo control se estima mediante el estimador de Kaplan-Meier, $\hat{S}_{0KM}(t)$, y fue obtenida a partir de un estudio anterior. Si se requiere detectar la razón de riesgo Δ, la hipótesis alterna de la función de sobrevida, asumiendo riesgos proporcionales, es

$$\hat{S}_1(t) = \left(\hat{S}_0(t)\right)^{1/\Delta}.$$

Para estimar el número esperado de eventos, con un periodo de acumulación (ingreso) a y de seguimiento f, se usa la media ponderada de sobrevida

$$\hat{S}(t) = p\hat{S}_0(t) + (1-p)\hat{S}_1(t),$$

donde p es la proporción de unidades asignadas aleatoriamente al grupo control.

Dada la función de sobrevida $\hat{S}(t)$, la integral (11.3.1) se puede evaluar numéricamente. La integral $\pi = 1 - \frac{1}{a}\int_f^{a+f} S(u,\theta)du$ puede escribirse como la suma de las áreas de los rectángulos bajo cada "paso" en los tiempos de falla entre a y $a+f$. Para hacer esto se consideran los tiempos al evento ordenados $t_{(1)}, \ldots, t_{(n)}$. La integral anterior se puede estimar como sigue:

$$\tilde{\pi} = \sum_{t_{(i)}: f < t_{(i)} \leq a+f} [S(a+f-t_{(i)}) \cdot (t_{(i)} - t_{(i-1)})]. \tag{11.6.1}$$

La probabilidad del evento puede ser computada con el siguiente código R.

```
library(survival)
S.km <- survfit(Surv(timeMonths, delta) ~ 1,
conf.type="log-log")
timesXe <- result.km$time
survXe <- result.km$surv
```

Luego se establece el tiempo de acumulación, a (ingreso de unidades), y el de seguimiento, f, y se selecciona la porción de tiempos al evento desde f hasta $a+f$, teniendo cuidado de incluir el tiempo f en el primer rectángulo:

```
accrual <- 12
followup <- 6
times.use <- c(followup, timesXe[{timesXe >= followup} &
{timesXe <= accrual + followup}])
surv.use <- summary(S.km, times=times.use)$surv
```

Finalmente, se usa la función "diff" para obtener el ancho de los rectángulos y la función "sum" completa la evaluación de $\pi = 1 - \frac{1}{a}\int_f^{a+f} S(u,\theta)du$.

```
> times.diff <- diff(c(times.use, accrual + followup))
> pi.rec <- 1 - (1/accrual)*sum(times.diff*surv.use)
> pi.rec
[1] 0.5365161
```

Así, el estimador de π mediante el "método del rectángulo" es $\pi_r \approx 0.5365$.

El número de unidades necesarias para el estudio es estimado mediante $n = e/\pi$, donde e es el número de eventos requerido.

Capítulo

doce

Modelamiento conjunto de datos longitudinales y de sobrevida

$E(Yi) = X_i\beta$

12.1. Introducción

En los estudios de seguimiento, generalmente en los prospectivos, se recolectan diferentes tipos de respuestas para cada unidad de una muestra. Estas pueden incluir varias respuestas medidas repetidamente en el tiempo (longitudinalmente) y el tiempo hasta que ocurre un evento de interés particular. Así, esta estructura la conforman los datos en medidas repetidas o longitudinales y los datos de tiempo para un evento o de sobrevida.

Las preguntas de interés frecuentemente conllevan a realizar análisis por separado de los datos longitudinales y de los datos de tiempo para el evento, respectivamente. Pero en muchas ocasiones, el interés de la investigación se dirige a indagar, explorar y confirmar la posible estructura, asociación o correlación entre los dos tipos de datos. Con frecuencia se presentan situaciones en las cuales se miden repetidamente, sobre una misma unidad, tanto la respuesta de marcadores como el tiempo hasta la ocurrencia de un evento. Se espera que los datos longitudinales o marcadores, junto con algunas covariables de la unidad, den cuenta del desarrollo del evento de interés.

El propósito de la medición de variables longitudinales es el seguimiento y pronóstico de la ocurrencia del evento de interés. Así, se mide sobre una misma unidad el tiempo que transcurre hasta la ocurrencia de un evento junto con algunas covariables. Dentro del conjunto de covariables, algunas pueden ser dependientes del tiempo o dinámicas, mientras que otras pueden ser no dependientes del tiempo.

En estos casos, como se ha advertido, se pueden agrupar los datos asociados a tres grupos de variables: el tiempo para el evento, las medidas repetidas (o longitudinales) y las covariables. Estas últimas pueden afectar alguno de los dos primeros grupos o ambos (datos de sobrevida o longitudinales). Varias aproximaciones han sido propuestas en la literatura para modelar, de forma separada, medidas repetidas y datos de sobrevida. Uno de los modelos usados con mayor frecuencia para el análisis de datos longitudinales es el modelo de efectos mixtos, como se puede consultar en Diggle *et al.* [23], Verbeke y Molenberghs [70, 71], Fitzmaurice *et al.* [26], Liu [51] o Díaz y León [20]. Los datos de tiempo para un evento (de sobrevida) han sido estudiados ampliamente a través de modelos de riesgo proporcional y de modelos acelerados en Kalbfleisch y Prentice [37], Klein [39], Collet [10], Smith [64], entre otros.

En la siguiente sección se resume la metodología para el análisis y modelamiento de los datos longitudinales de manera separada de los datos de tiempo para un evento o de sobrevida, pues el modelamiento de estos últimos datos ha sido el objeto de este texto. En la última sección se presenta el

modelamiento conjunto, que integra el modelamiento de los datos en medidas repetidas o longitudinales y los de tiempo para la ocurrencia de un evento (sobrevida).

12.2. Modelamiento de datos longitudinales

Aunque se ha argumentado que no siempre es adecuado desarrollar únicamente un análisis, de manera separada, de cada uno de estos tres grupos de variables, en esta parte se presentan algunas técnicas para el modelamiento y análisis estadístico de los datos en medidas repetidas o longitudinales.

Suponga un estudio en el que se tienen m unidades, las cuales son seguidas durante un periodo de tiempo $[0, \tau)$. En general, los datos asociados con estos estudios son desbalanceados, en el sentido de que el número de medidas registradas para las unidades no es el mismo ni se hace en los mismos tiempos. Así, cada unidad i está asociada con un vector Y_i de respuestas $\{Y_{ij} : \ j = 1, \ldots, n_i\}$; explícitamente, $Y_i = (Y_{i1}, \ldots, Y_{in_i})$.

Los datos longitudinales tienen, por lo menos, tres componentes de variación aleatoria: (i) variabilidad entre unidades (*efectos aleatorios*), (ii) correlación serial y (iii) errores de medición. Una tarea importante con el análisis de datos longitudinales es la formulación de un modelo apropiado para describir la evolución de la variable respuesta sobre el tiempo en función de unas covariables. Estos modelos deben incluir o considerar los componentes de variación aleatoria mencionados.

Modelo lineal mixto

Para ajustar un modelo a los datos longitudinales, los modelos lineales mixtos son usados con frecuencia para variables continuas o discretas [70, 71, 23, 54, 26]. En general, un modelo lineal de efectos mixtos es cualquier modelo que satisface

$$\begin{cases} Y_i = X_i\beta + Z_i b_i + \epsilon_i \\ b_i \sim N(0, D), \\ \epsilon_i \sim N(0, R_i), \\ b_1, \ldots, b_m, \epsilon_1, \ldots, \epsilon_m \ independendientes, \end{cases} \tag{12.2.1}$$

donde

- Y_i es el vector de respuestas n_i-dimensional para la unidad $i = 1, \ldots, m$,

- X_i es una $(n_i \times p)$ "matriz de diseño", la cual caracteriza la parte sistemática de la respuesta, es decir, depende de las covariables y del tiempo,
- β es un vector p-dimensional que contiene los efectos fijos,
- Z_i es una $(n_i \times q)$ "matriz de diseño" que caracteriza la variación aleatoria en la respuesta, la cual es atribuible a la variabilidad entre de las unidades,
- b_i es un vector q-dimensional que contiene los efectos aleatorios que complementan la caracterización de la variación entre las unidades,
- ϵ_i es un vector n_i-dimensional de los errores dentro de las unidades que caracteriza la variación debida a la manera en que las respuestas son medidas sobre la unidad i,
- D es una matriz de covarianzas $(q \times q)$ con elemento (i, j) $d_{ij} = d_{ji}$, y
- R_i es un matriz de covarianzas $(n_i \times n_i)$.

En el análisis de datos longitudinales se debe tener en cuenta, y hay que involucrar en el modelo, la dependencia entre las múltiples medidas registradas sobre una misma unidad. La dependencia o correlación se asume incluida en la variable no observable Z_i.

El modelo lineal general mixto (12.2.1) implica el modelo marginal

$$Y_i \sim N(X_i\beta, \Sigma_i), \text{ para } \Sigma_i = Z_i D Z_i' + R_i. \tag{12.2.2}$$

Una vez que el modelo ha sido postulado, el problema es la estimación de los parámetros que lo conforman. Los métodos de *máxima verosimilitud* y *máxima verosimiltud restringida (REML)* se pueden usar para estimar los parámetros que caracterizan la "media" o la parte sistemática del modelo, β, y los que caracterizan la "variación" o la parte aleatoria del modelo, junto con los parámetros que conforman las matrices de covarianzas D y R_i.

Un objetivo del análisis es caracterizar el comportamiento o trayectoria de cada unidad o individuo. Así, se puede afirmar que los modelos lineales de efectos mixtos son modelos para unidades o sujetos específicos, en el sentido de que se ajustan a un "modelo de regresión" para la i-ésima unidad, el cual tiene "media" $X_i\beta + Z_i b_i$. Esto implica estimar β y "estimar" b_i. No obstante, ya que b_i no es una constante fija como β, el interés se dirige a la "predicción" de la cantidad aleatoria b_i. Así, para caracterizar el comportamiento de una unidad dentro del modelo, se desarrolla un método para la predicción de b_i.

Se demuestra en McCulloch y Searle [54, p. 168] que la esperanza condicional de b_i, dados los datos de Y_i, es

$$\mathscr{E}(b_i|y_i) = DZ_i'\Sigma_i^{-1}(Y_i - X_i\beta). \tag{12.2.3}$$

Un caso especial se presenta cuando D y R_i son conocidas y $\hat{\beta}$ es el estimador de mínimos cuadrados ponderados (WLS). En este caso,

$$DZ_i'\Sigma_i^{-1}(Y_i - X_i\hat{\beta}) \tag{12.2.4}$$

es el *mejor predictor lineal insesgado* (BLUP) para b_i.

Una aproximación para el BLUP de b_i es

$$\hat{b_i} = \hat{D}Z_i'\hat{\Sigma_i^{-1}}(Y_i - X_i\hat{\beta}). \tag{12.2.5}$$

Este "estimador" es frecuentemente referido como el *estimador empírico de Bayes* para b_i.

Modelo lineal generalizado

Hasta aquí, el modelo ha sido estructurado para situaciones donde:

i) La respuesta es continua y tiene distribución normal.

ii) La relación entre la media de la variable respuesta, Y, y las covariables (incluido el tiempo), X, se establece mediante un modelo lineal en los parámetros, es decir, $\mathscr{E}(Y_i) = X_i'\beta$.

iii) La varianza de Y_i es constante e independiente de los X; $var(Y_i) = \sigma^2$.

Un modelo alternativo para tales datos es el *modelo lineal generalizado* y el libro pionero de estos modelos es el de McCullagh y Nelder [53]. Un modelo lineal generalizado es una regresión para la variable respuesta Y_i con la siguiente estructura:

i) La distribución de probabilidad para Y_i se asume que pertenece a la familia exponencial.

ii) La media de Y_i se considera de la forma

$$\mathscr{E}(Y_i) = f(X_i'\beta). \tag{12.2.6}$$

La función $f(\cdot)$ es monótona y diferenciable. Esto significa que existe una única función g, llamada función inversa de f, por lo que, de esta manera, se puede escribir el modelo en la forma equivalente a (12.2.6):

$$g\{\mathscr{E}(Y_i)\} = X_i'\beta. \tag{12.2.7}$$

iii) La varianza de las Y_i es de la forma

$$var(Y_i) = \phi V\{\mathscr{E}(Y_i)\},$$

donde la función V depende de la distribución particular de las Y_i y ϕ puede ser igual a una constante conocida.

Una aproximación natural para estimar β en estos modelos es mediante el principio de máxima verosimilitud. Así, el estimador de máxima verosimilitud para β corresponde a la solución del sistema de ecuaciones

$$\sum_{i=1}^{m} \frac{1}{V\{f(X_i'\beta)\}}\{Y_i - f(X_i'\beta)\}f'(X_i'\beta)X_i = 0. \tag{12.2.8}$$

La extensión de los modelos lineales clásicos al modelamiento de datos longitudinales se facilita al considerar la distribución normal multivariada de estos datos; para ello se puede consultar a Díaz y Morales [21]. Por analogía, es natural esperar que cuando los elementos del vector de datos Y_i sean conteos, respuestas binarias o respuestas positivas, en estos casos se podría optar por los modelos lineales generalizados con la inclusión de la correlación debida a las mediciones repetidas sobre una misma unidad. No obstante, este no es el procedimiento adecuado y la misma generalización no es fácil ni pertinente. Así, tratar de usar extensiones multivariadas de los modelos lineales generalizados para el caso de vectores de datos longitudinales resulta demasiado complejo para producir modelos útiles en escenarios reales.

Otra complicación es el modelamiento de la media poblacional y de las unidades específicas. La media poblacional es considerada como la respuesta media, a través de todas las unidades, en cada uno de los puntos en el tiempo. La media poblacional describe la relación y evolución de las respuestas en diferentes tiempos. El modelo describe la respuesta media en cualquier tiempo para una unidad como una función de parámetros fijos y posiblemente algunas covariables.

El artículo publicado por Liang y Zegel [46] resuelve algunos de los inconvenientes planteados. Lo que proponen los autores es olvidarse de tratar de modelar la media global de la distribución multivariada de probabilidad del vector de datos. En su lugar, la idea es tan solo modelar la respuesta media y la matriz de covarianzas del vector de datos como en el caso de la distribución normal. De otra manera, se trata de considerar los modelos lineales generalizados para observaciones individuales como punto de partida. Esta es la estrategia de modelamiento:

- La *respuesta media* del vector de datos Y_i es modelada como una función del tiempo, las covariables X y los parámetros β, mediante

modelos lineales generalizados que representan la respuesta media de cada unidad Y_i.

- La *varianza* de cada unidad Y_i es modelada por una función de la media de acuerdo con la distribución de los datos. Estos modelos son frecuentemente modificados o transformados para permitir la inclusión de la variación dentro y entre unidades mediante la adición de parámetros de dispersión ϕ.

- La *correlación* entre observaciones sobre la misma unidad (elementos de Y_i) se representa mediante una adecuada selección de la estructura de correlación. La matriz de correlación seleccionada es referida como la *matriz trabajo*.

Con estas consideraciones, se tiene el siguiente modelo estadístico para modelar la media y la matriz de covarianzas del vector de datos Y_i, el cual tiene las observaciones Y_{ij}, $j = 1, \ldots, n_i$ sobre la i-ésima unidad:

- La respuesta media, μ, de Y_{ij} es modelada mediante una función apropiada g, denominada *función de enlace*, la cual asigna a la media un predictor lineal de las covariables, es decir, $g(\mu) = X'_{ij}\beta$.

- La varianza es entonces modelada por alguna función $V(\cdot)$ de la respuesta media y por un parámetro de dispersión ϕ, lo cual define una matriz $T_i^{1/2} = diag(\{V(\mu_{ij})\}^{1/2})$. La correlación es modelada mediante una matriz de "trabajo" Γ_i.

Esto se resume en las siguientes expresiones:

$$\mathscr{E}(Y_i) = \begin{pmatrix} h(X'_{i1}\beta) \\ h(X'_{i2}\beta) \\ \vdots \\ h(X'_{in_i}\beta) \end{pmatrix} = h_i(\beta), \; var(Y_i) = \phi T_i^{1/2}\Gamma_i T_i^{1/2} = \Sigma_i = \phi\Lambda_i. \tag{12.2.9}$$

A partir de lo anterior, se puede decir que el conocimiento de la media y la matriz de covarianzas no es suficiente para la especificación completa del modelo probabilístico multivariado asociado con las observaciones del vector de datos longitudinales. Una salida a esta situación es dada por Liang y Zeger [46], quienes hacen caso omiso del modelo probabilístico que genera los datos longitudinales y parten de la idea de que, para el modelamiento de los datos, tan solo se requiere la media y la matriz de covarianzas. Esta es la metodología que se presenta a continuación.

Ecuaciones de estimación generalizadas

La idea es generalizar y extender las ecuaciones de estimación de verosimilitud usuales para un modelo lineal generalizado de variable respuesta univariada que incorpore la correlación debida a las múltiples observaciones sobre una misma unidad. Al final de la sección anterior, se resaltó que la especificación del modelo para la media y la matriz de covarianzas de un vector de datos se presenta en la forma (12.2.9). Sin embargo, esto no es suficiente para especificar, de manera completa, la distribución de probabilidad multivariada.

Para ilustrar esto se muestran dos situaciones. La primera es el caso de la distribución normal para un modelo lineal de la media. El modelo es

$$\mathscr{E}(Y_i) = X_i\beta\,,\ var(Y_i) = \Sigma_i$$

para una selección apropiada de la matriz de covarianzas Σ_i. Se asume que los Y_i siguen una distribución normal multivariada. El estimador para β es

$$\hat{\beta} = \Big(\sum_{i=1}^{m} X_i'\hat{\Sigma}_i{}^{-1}X_i\Big)^{-1}\sum_{i=1}^{m} X_i'\hat{\Sigma}_i{}^{-1}Y_i.$$

Esta ecuación puede escribirse en la forma:

$$\sum_{i=1}^{m} X_i'\hat{\Sigma}_i{}^{-1}(Y_i - X_i\hat{\beta}) = 0, \tag{12.2.10}$$

la cual corresponde a las ecuaciones para modelos lineales generalizados. El estimador para β se obtiene como solución de (12.2.8); es decir,

$$\sum_{i=1}^{m} \frac{1}{V\{f(X_i'\beta)\}}\{Y_i - f(X_i'\beta)\}f'(X_i'\beta)X_i = 0. \tag{12.2.11}$$

El supuesto de normalidad no siempre es seguido por un conjunto de datos. Por ejemplo, si la respuesta está en la forma de conteos o es binaria, resulta claro que la distribución normal no es apropiada.

Al comparar (12.2.10) y (12.2.11), se observan algunas coincidencias, como que: las ecuaciones son funciones lineales de los desvíos de las observaciones respecto a la media asumida, ponderadas de acuerdo con su matriz de covarianzas. A partir de estas observaciones, una aproximación natural para ajustar el modelo (12.2.9) consiste en resolver un sistema de ecuaciones de estimación que consiste de p ecuaciones para β, las cuales son funciones lineales de los desvíos

$$Y_i - f_i(\beta).$$

Las ponderaciones de estos desvíos son de la misma forma contenida en (12.2.10) y (12.2.11), mediante la inversa de la matriz de covarianza Σ_i del vector de datos.

Estos resultados permiten considerar las siguientes ecuaciones y resolver para los parámetros β

$$\sum_{i=1}^{m} \Delta_i' \Lambda_i^{-1} \{Y_i - f_i(\beta)\} = 0, \tag{12.2.12}$$

donde Δ_i es la $(n_i \times p)$ matriz cuya elemento (j, s) es la derivada de $f(X_{ij}'\beta)$ respecto al s-ésimo elemento de β, y Λ_i es la matriz Σ_i en (12.2.9).

Un sistema de ecuaciones de la forma (12.2.12), cuya solución corresponde a una estimación de los parámetros β en un modelo para la media, es referido como *ecuaciones de estimación generalizada* (*GEE*, por su sigla en inglés).

El modelo expresado en (12.2.1) puede ser extendido al caso de datos no normales. La idea es modelar las trayectorias individuales, donde la "media" en el tiempo t_{ij} sobre todas las observaciones es representada nuevamente por un *modelo lineal generalizado* para una trayectoria individual específica. Esto es un modelo lineal mixto generalizado [54, 26], en el que se permite que los parámetros dependan de *efectos aleatorios*. La esperanza condicional de Y_i, dado un vector de efectos aleatorios b_i para la unidad i, puede ser considerada como la respuesta "media" para una unidad particular. Una representación general de tales modelos es como sigue:

$$\mathscr{E}(Y_{ij}|b_i) = g(X_j'\beta_i), \tag{12.2.13}$$

donde $g(\cdot)$ es una función apropiada, β_i es un parámetro de una unidad específica, el cual en el caso más general puede expresarse como

$$\beta_i = A_i\beta + B_i b_i. \tag{12.2.14}$$

Aquí, β es el parámetro que describe el valor "típico" de β_i's a través de todas las unidades con matriz de covariables A_i (es decir, todos las unidades bajo un tratamiento particular). b_i es un efecto aleatorio, el cual se asume que es generado por una distribución con media 0 y matriz de covarianzas D. Además, se asume que, en el nivel individual, los datos en Y_i siguen una distribución que pertenece a la familia exponencial escalada.

El modelo definido en (12.2.13) y (12.2.14) es referido como *modelo lineal mixto generalizado*. Con estos modelos, la respuesta media es modelada como una combinación de las características poblacionales (efectos fijos), β, las cuales se asumen compartidas por todas las unidades, y los efectos

individuales específicos (efectos aleatorios), que son únicos para una unidad o individuo particular.

12.3. Modelamiento conjunto en datos longitudinales y de sobrevida

En estudios longitudinales donde hay medidas repetidas sobre una respuesta continua, una observación sobre un tiempo hasta el evento posiblemente censurado, junto con la información de algunas covariables que son tomadas sobre cada unidad, son comunes en investigación médica, en ingeniería o en ciencias sociales, entre otras. El interés frecuentemente se dirige a explorar, modelar y confirmar las posibles interrelaciones entre estos tres conjuntos de variables (medidas repetidas, tiempo al evento y covariables). Por ser de interés didáctico, se resume el artículo de unos de los pioneros del modelamiento conjunto de datos longitudinales y de sobrevida, Tsiatis y Davidian [67].

Un ejemplo familiar son los ensayos clínicos sobre VIH, donde las covariables, incluida la asignación de tratamiento, la información demográfica y las características fisiológicas, se registran al inicio del estudio. Además, se registran medidas inmunológicas y del estado virológico, como el recuento de células CD4 y el número de copias del ARN viral que se toman en visitas posteriores de los pacientes a la clínica. El tiempo para desarrollar el SIDA o la muerte también se registra para cada participante del estudio, aunque algunos sujetos pueden retirarse de manera temprana o no experimentar el evento al momento del cierre del estudio. El estudio pudo haber sido diseñado para abordar la pregunta principal del efecto del tratamiento sobre el tiempo hasta el evento, pero los objetivos subsiguientes pueden ser (i) comprender los patrones de cambio de CD4 o carga viral o (ii) caracterizar la relación entre las características de CD4 o perfiles de carga viral y el tiempo hasta el desarrollo de la enfermedad o la muerte. De manera similar, en estudios de cáncer de próstata, se pueden obtener las mediciones repetidas del antígeno prostático específico (en inglés, PSA) en pacientes que siguen un tratamiento, junto con el tiempo hasta la recurrencia de la enfermedad. De nuevo, puede ser de interés caracterizar (i) el patrón de variabilidad del PSA dentro de cada sujeto y (ii) la asociación entre las características del proceso longitudinal del PSA y la recurrencia del cáncer.

Los objetivos (i) y (ii) pueden hacerse más precisos al considerar que los datos para cada sujeto $i = 1, \ldots, n$ son $\{T_i,\ Z_i,\ X_i(u),\ u \geq 0\}$, donde T_i es el tiempo al evento, Z_i es un vector de covariables de referencia (al tiempo

0), y $\{X_i(u),\ u \geq 0\}$ es la trayectoria de la respuesta longitudinal para todos los tiempos $u \geq 0$. El objetivo (i), hacer explícitos los patrones de cambio de la respuesta longitudinal (por ejemplo, CD4 o PSA), puede implicar, por ejemplo, estimar aspectos del comportamiento longitudinal promedio y su asociación con covariables, como $E\{X_i(u)|Z_i\}$ y $Var\{X_i(u)|Z_i\}$. Un esquema de trabajo usual para abordar el objetivo (ii), que es caracterizar las asociaciones entre los procesos longitudinales, el tiempo hasta el evento y las covariables, consiste en representar la relación entre T_i, $X_i(u)$ y Z_i mediante un modelo de riesgos proporcionales

$$h(_i(u)) = h_0(u)\exp\{\gamma X_i(u) + \eta' Z_i\}, \tag{12.3.1}$$

donde $X_i^H(u) = \{X_i(t), 0 \leq t < u\}$ es el historial del proceso longitudinal hasta al tiempo u, como en Kalbfleisch y Prentice [37, sec. 6.3], y $X_i(u)$ debería ser continua a izquierda; en lo sucesivo, se considera solo X_i(u) continuo. En (12.3.1), por definición, se considera que el riesgo depende linealmente de la historia a través del valor actual $X_i(u)$, aunque otras especificaciones y formas de relación son posibles. Aquí, $X_i(u)$ puede considerarse como una covariable tiempo dependiente en (12.3.1), y la evaluación formal de la asociación implica la estimación de γ y η. El interés del primer autor en este problema surgió de su trabajo con el Grupo de Ensayos Clínicos sobre el SIDA (ACTG) en la década de 1990, cuando era muy debatida la pregunta de si el recuento longitudinal de células CD4 es un "marcador sustituto" del tiempo hasta el desarrollo del SIDA o la muerte. Es decir, ¿puede el punto final de supervivencia, generalmente a largo plazo, ser reemplazado por medidas longitudinales de las células CD4 a corto plazo para evaluar la eficacia del tratamiento? Prentice [59] estableció las condiciones de subrogación:

I) El tratamiento debe tener un efecto en el tiempo hasta el evento.

II) El tratamiento debe tener efecto en el marcador.

III) El efecto del tratamiento debe manifestarse a través del marcador, es decir, el riesgo del evento, dada una trayectoria específica del marcador, debe ser independiente del tratamiento.

Un marcador sustituto exitoso debe ser capaz de reducir el tiempo de seguimiento o reducir el número de pacientes necesarios para establecer un determinado efecto de un tratamiento. En el libro de Burzykowski, Molenberghs y Buyse [7] se presenta el concepto y validación de los biomarcadores

y se muestra la combinación de los datos longitudinales y de tiempo para un evento en los modelos conjuntos.

Las principales justificaciones para el modelamiento conjunto de datos longitudinales y tiempo para el evento son las siguientes:

- Las covariables dependientes del tiempo se registran intermitentemente sobre cada unidad y están posiblemente sujetas a error.
- Los datos longitudinales puede resultar truncados debido a la ocurrencia del evento, situación que posiblemente induce datos longitudinales faltantes de tipo informativo.
- La posible imputación de algunos datos en el tiempo que se presenta el evento, para considerarlos en un modelo de riesgos proporcionales con covariables dependientes del tiempo, induciría un sesgo en los coeficientes de regresión estimados.

El modelamiento conjunto de datos longitudinales y tiempo para el evento ha sido propuestos como una respuesta a las dificultades anteriores. El modelamiento conjunto tiene dos objetivos principales:

I) Explorar y determinar dentro de cada unidad el modelo de cambio. En estos modelos la atención principal está en la tendencia serial.

II) Caracterizar la relación entre las medidas repetidas y el tiempo para el evento. Aquí, el objetivo de los modelos conjuntos es la descripción y la significación del efecto de las variables longitudinales sobre la respuesta de sobrevida.

A continuación se presenta, en orden cronológico, una revisión de los artículos pioneros en el modelamiento conjunto de datos longitudinales y tiempo para el evento.

12.3.1. DeGruttola y Tu (1994)

El artículo pionero es de Degruttola y Tu [18]. Ellos modelan la regresión de las células CD4 y su asociación entre diferentes características del desarrollo del VIH y el tiempo de sobrevida. También extienden el modelo general de efectos aleatorios al análisis de datos longitudinales en presencia de censura informativa y proponen un modelo en el cual ambas variables, los conteos repetidos de células CD4 y el tiempo de sobrevida, son conjuntamente incluidas en el modelo. Además de la estimación, el método permite

describir la relación entre las características del desarrollo individual de la enfermedad y el tiempo de sobrevida.

Sea y_i un vector $n_i \times 1$, cuyos elementos, y_{ij}, son los valores del marcador para el i-ésimo individuo en la j-ésima ocasión para $i = 1, \ldots, m$ y $j = 1, \ldots, n_i$. Sea $\tau_i = \text{mín}\{x_i, c_i\}$, donde x_i y c_i denotan el tiempo de sobrevida y la censura para el i-ésimo individuo, respectivamente. La evolución del marcador, y_i, es modelada mediante una variable aleatoria en un modelo de la forma

$$y_i = T_i\alpha + Z_i b_i + \epsilon_i, \ i = 1, \ldots, m. \tag{12.3.2}$$

Aquí, α es vector $p \times 1$ de parámetros desconocidos; T_i es una matriz de diseño de rango completo $n_i \times p$; $b_i \sim iidN(0; D)$ que denota un vector $k \times 1$ de los efectos desconocidos de los individuos; Z_i es una matriz $n_i \times k$ de diseño conocida; $\epsilon_i \sim_{iid} N(0, \sigma^2 I_{n_i})$ es un vector de residuales, e I_{n_i} es una $n_i \times n_i$ matriz identidad.

Si el tiempo de sobrevida depende de la verdadera pero no observada variación de células CD4, esto es, $\phi(x|b, y) \neq \phi(x|y)$, así, la relación entre x y b debe ser modelada explícitamente. Para ello, se consideran: el momento de la medida, t_j; el conteo observado de células CD4, y_j, en el tiempo t_j; el tiempo para la muerte, x, o la censura, c; y el indicador de censura ρ. Se asume que la censura es no informativa. También se asumen que la probabilidad condicional de muerte, dado el efecto aleatorio b, es independiente del proceso del marcador. Con estas consideraciones, la verosimilitud para cualquier individuo i es

$$L = \phi(y_1|b) \times \prod_{j=2}^{M} \left\{ [\phi(y_j|b) g(t_j|y_1, \ldots, y_{j-1}; t_j < x)]^{I(t_j<x)} [\phi(x|b)]^{\rho} [1 - \Phi(c|b)]^{1-\rho} \right\}, \tag{12.3.3}$$

donde M es el número máximo de posibles ocasiones de medición, y

$$g(t_j|y_1, \ldots, y_{j-1}; t_j < x)$$

es la probabilidad condicional de que el tiempo de sobrevida de un individuo sea t_j.

Para modelar el tiempo de sobrevida x, considere

$$x_i = w_i^T \zeta + \lambda^T b_i + r_i, \tag{12.3.4}$$

donde ζ es un vector de $q \times 1$ parámetros desconocidos; w_i es una matriz de diseño $q \times 1$; λ es un vector de $k \times 1$ parámetros desconocidos; y $r_i \sim_{iid} N(0, s^2)$ es un residual.

Note que los modelos (12.3.2) y (12.3.4) tienen el parámetro común b_i, de otra manera, la sobrevida depende del parámetro que describe la verdadera pero no observada trayectoria del conteo de linfocitos CD4, aunque puede también depender de otras covariables w_i. Estas covariables pueden incluir ambas, las longitudinales y de las sobrevida.

El logaritmo de la función de verosimilitud conjunta puede escribirse mediante la integración de (12.3.3) sobre el rango no observable b y la suma sobre los individuos, así:

$$L_{obs} = \sum_{i=1}^{N^0} \log\left[\int_{b_i} \phi(y_i|b_i, \alpha, \sigma^2)\phi(x_i|D)\phi(x_i|b_i, \zeta, s^2)db_i\right] + \sum_{i=1}^{N^c} \log\left[\int_{b_i} \phi(y_i|b_i, \alpha, \sigma^2)\phi(x_i|D)(1 - \Phi(c_i|b_i, \zeta, s^2))db_i\right], \quad (12.3.5)$$

donde N^o y N^c denotan el número de individuos muertos y censurados, respectivamente. Para estimar los parámetros del modelo, se hace uso de una adaptación del algoritmo EM [19].

12.3.2. Tsiatis, DeGruttola y Wulfsohn (1995)

Estos autores usan el modelo de regresión de riesgos proporcionales de Cox para estudiar la relación entre el conteo de células CD4, como una covariable dependiente del tiempo, y el tiempo de sobrevida. Dado que el conteo de células se mide solo periódicamente, con un error de medición substancial y una variación biológica, los métodos clásicos para estimar los parámetros en el modelo de Cox no son apropiados. En su lugar, los autores proponen una aproximación en *dos etapas*. En la primera, el conteo longitudinal de CD4 se modela mediante el modelo de componente aleatorio para medidas repetidas. En la segunda etapa, con el propósito de estimar los parámetros del modelo de Cox, se asume que los datos son los de sobrevida o los de tiempo para el evento. Aquí es necesario tener un conocimiento completo de la historia de las covariables para todas las unidades, las cuales son obtenidas en la primer etapa y luego se reemplazan en la verosimilitud parcial para el modelo de Cox con covariables dependientes del tiempo. Los parámetros se estiman desde la maximización de la verosimilitud parcial.

Para la construcción del modelo, considere a $Z^*(t)$ un valor hipotético de conteo de células CD4 en el tiempo t, donde este conteo se asume sin

error de medida. Considere que $\bar{Z}^*(t)$ denota la historia hasta el tiempo t, $\{Z^*(u) : u \leq t\}$. El objetivo principal es estimar la relación entre el tiempo de sobrevida y el proceso de conteo verdadero de células. Sea T el tiempo de sobrevida de un individuo, la tasa de riesgo en el momento t, como una función de la historia de la covariable longitudinal hasta el tiempo t, es definida como

$$\lambda(t|\bar{Z}^*(t)) = \lim_{h\to 0} \frac{1}{h}\{pr(t \leq T < t+h|T \geq t, \bar{Z}^*(t))\}. \tag{12.3.6}$$

Los autores asumen que los datos de sobrevida están sujetos a censura no informativa a la derecha. Sea $X = \text{mín}\{T, C\}$, donde C corresponde a un tiempo de censura potencial, y el indicador de falla es Δ, el cual es igual a 1 si el individuo es observado para fallar al tiempo $(T \leq C)$ y es 0 en otro caso.

Como la censura es no informativa, la tasa de riesgo expresada en (12.3.6) es igual a $\lambda(t|\bar{Z}^*(t))$. Así, el modelo de riesgo proporcional de Cox relaciona el riesgo con las covariables dependientes del tiempo, es decir,

$$\lambda(t|\bar{Z}^*(t)) = \lambda_0(t) f(\bar{Z}^*(t), \beta), \tag{12.3.7}$$

donde $f(\bar{Z}^*(t), \beta)$ es una función de la historia de las covariables especificadas para un parámetro β (posiblemente un vector).

El conocimiento de $Z^*(t)$, para todos los valores de $t \leq X$, generalmente no es registrado. Así, debe ser considerado el efecto del error de medición, esto es,

$$Z(t) = Z^*(t) + e(t).$$

Se asume que $e(t)$ es un ruido aleatorio, tal que $\mathscr{E}\{e(t)\} = 0, var\{e(t)\} = \sigma^2$ y $cov\{e(s), e(t)\} = 0, s \neq t$.

Se nota por $\bar{\mathbf{Z}}(t) = \{Z(t_1), \ldots, Z(t_j);\ t_j \leq t\}$ la historia de los conteos observados de CD4 hasta el tiempo t. Así, el riesgo observable es $\lambda(t|\bar{\mathbf{Z}}(t))$, en lugar de $\lambda(t|\bar{\mathbf{Z}}^*(t))$. Por propiedades de la probabilidad condicional:

$$\lambda(t|\bar{Z}(t)) = \int \lambda(t|\bar{Z}(t), \bar{Z}^*(t)) dP(\bar{Z}^*(t)|\bar{Z}(t), X \geq t).$$

Si se asume que no hay errores de medición y que los tiempos de las visitas son predecibles, a partir de (12.3.7), se tiene

$$\lambda(t|\bar{Z}(t), \bar{Z}^*(t)) = \lambda(t|\bar{Z}^*(t)) = \lambda_0(t) f(\bar{Z}^*(t), \beta),$$

así,

$$\lambda(t|\bar{Z}(t)) = \lambda_0(t)\mathscr{E}[f(\bar{Z}^*(t), \beta)|Z(t_1), \ldots, Z(t_j), X \geq t]. \tag{12.3.8}$$

La esperanza condicional en (12.3.8) se nota por $\mathscr{E}(t, \beta)$. Si $\mathscr{E}(t, \beta)$ es conocido, entonces β puede ser estimado por la maximización de la verosimilitud parcial,

$$\prod_{i=1}^{m}\left[\mathscr{E}_i(X_i, \beta)\Big/\sum_{j=1}^{m}\mathscr{E}_j(X_j, \beta)\right]^{\Delta_i}. \tag{12.3.9}$$

Para ajustar el modelo, los autores proponen un procedimiento bietápico. Primero, para el verdadero conteo de CD4, asumen una curva de crecimiento para un modelo de componentes aleatorios con errores distribuidos normalmente y emplean el algoritmo EM, modificado por Laird y Ware [41] para la estimación de los parámetros. Luego sustituyen estas estimaciones (llamadas estimadores de Bayes empíricos) en el modelo de riesgos proporcionales para obtener estimaciones de los parámetros del modelo de sobrevida mediante verosimilitud parcial. Este procedimiento se resume así:

- Primera etapa: los conteos log CD4 siguen una curva de crecimiento lineal de la forma

$$Z_i(u) = Z_i^*(u) + e_i(u), \quad u \le t, \text{ donde } Z_i^*(u) = \theta_{0i} + \theta_{1i}u. \tag{12.3.10}$$

- Segunda etapa: la función de riesgo puede ser modelada para el conteo histórico de CD4 como

$$\lambda_i(t) = \lambda_0(t)\exp\{\beta_1(\theta_{0i} + \theta_{1i}t) + \beta_2\theta_{1i}\}. \tag{12.3.11}$$

12.3.3. Wulfsohn y Tsiatis (1997)

Los autores consideran, nuevamente, la relación entre un proceso longitudinal y un proceso de tiempo para un evento. También abordan el problema de estimar los parámetros del modelo de Cox cuando la covariable longitudinal es medida en forma intermitente y con errores de medición. Los estimadores de los parámetros se obtienen mediante la maximización de la verosimilitud conjunta para el proceso de covariable longitudinal y el proceso de tiempo de falla. Los parámetros son estimados por medio del algoritmo de la esperanza máxima (procedimiento EM).

Los datos observados para cada individuo son de la forma

$$(X_i, \Delta_i, Z_i, t_i),$$

donde $X_i = \text{mín}\{T_i, C_i\}$, con T_i siendo tiempo de sobrevida para el i-ésimo individuo y C_i siendo un tiempo de censura potencial para el i-ésimo individuo. Δ_i es el indicador de falla, el cual es igual a 1 si la falla es observada

(es decir, si $T_i \leq C_i$) y es igual a 0 en otro caso. Se asume que la censura es independiente del tiempo de sobrevida y de la covariable longitudinal. $t_i = (t_{ij} : t_{ij} \leq X_i)$ son los tiempos de observación para el i-ésimo individuo, para $i = 1, \ldots, m$ y $j = 1, \ldots, n_i$. $Z_i = (Z_{ij} : t_{ij} \leq X_i)$ son los valores de las covariables (posiblemente transformadas) sobre el i-ésimo individuo, los cuales son medidos en los tiempos t_i.

Los autores modelan la covariable longitudinal con una curva de crecimiento lineal con efectos aleatorios (intercepto y pendiente), de esta forma:

$$Z_{ij} = \theta_{0i} + \theta_{1i} t_{1j} + e_{ij},$$

donde e_{ij} se distribuye *iid* según $N(0, \sigma_e^2)$, $cov(e_{ij}, e_{ij'}) = 0$ (con $j \neq j'$). Tal error es independiente del intercepto y la pendiente aleatorios, θ_{0i} y θ_{1i}. Los autores asumen que las pendientes e interceptos individuales se distribuyen de acuerdo con una normal bivariada, esto es:

$$\begin{pmatrix} \theta_{0i} \\ \theta_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}, \quad \begin{pmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{01} & \sigma_{11} \end{pmatrix} \right).$$

El riesgo de falla es modelado mediante el modelo de Cox. En este modelo el riesgo depende de la covariable longitudinal a través de su valor actual. Se asume que el verdadero valor de la covariable longitudinal está dado por el modelo de curva de crecimiento con efectos aleatorios; esto es:

$$\lambda(t|\theta_i, Z_i, t_i) = \lambda_0(t) \exp\{\beta(\theta_{0i} + \theta_{1i} t)\}.$$

Una observación para cada individuo tiene la forma $(X_i, \Delta_i, Z_i, t_i)$. Los efectos aleatorios, θ_i, en el modelo de curva de crecimiento no son observables. La función de verosimilitud para estos datos está dada por

$$\prod_{i=1}^{m} \left[\int_{-\infty}^{\infty} \left\{ \prod_{j=1}^{n_i} f(z_{ij}|\theta_i, \sigma_e^2) \right\} f(\theta_i|\theta, V) f(X_i, \Delta_i|\theta_i, \lambda_0, \beta) d\theta_i \right], \tag{12.3.12}$$

donde

$$f(z_{ij}|\theta_i, \sigma_e^2) = (2\pi\sigma_e^2)^{-1/2} \exp\{-(z_{ij} - \theta_{0i} - \theta_{1i} t_{ij})^2\},$$
$$f(\theta_i|\theta, V) = (2\pi|V|)^{-1/2} \exp\{-(\theta_i - \theta)' V^{-1} (\theta_i - \theta)\},$$

y

$$f(X_i, \Delta_i|\theta_i, \lambda_0, \beta)$$
$$= [\lambda_0(X_i) \exp\{\beta(\theta_{0i} + \theta_{1i} X_i)\}]^{\Delta_i} \exp\left[-\int_0^{X_i} \lambda_0(u) \exp\{\beta(\theta_{0i} + \theta_{1i} u)\} du \right]. \tag{12.3.13}$$

La estimación de los parámetros θ, V, σ_e^2 y β se hace mediante máxima verosimilitud paramétrica y $\lambda_0(u)$ se estima de la máxima verosimilitud no paramétrica; se emplea el algoritmo EM de Dempster, Lair y Rudin [19] para determinar los estimadores de estos parámetros.

12.3.4. Henderson, Diggle y Dobson (2000)

Los autores desarrollan una metodología flexible para modelar de manera combinada los datos longitudinales y los datos históricos de un evento. Para la covariable longitudinal (dependiente del tiempo), ellos argumentan que en cualquier modelo general se deberían incorporar efectos fijos, efectos aleatorios, correlación serial y los errores de medición, mientras que la historia del evento, o análisis de sobrevida, debería estar basado en un modelo de riesgos proporcionales semiparamétrico con o sin término de fragilidad. Los métodos se ilustran mediante resultados de un ensayo clínico en el tratamiento de la esquizofrenia.

Se asume que la estructura de los datos para una unidad o individuo es una sucesión de medidas $\{Y_{ij} : j = 1, \ldots, n_i\}$ en los tiempos $\{t_{ij} : j = 1, \ldots, n_i\}$, con $i = 1, \ldots, m$, y un proceso de conteo $\{N_i(u) : 0 \leq u \leq \tau\}$, el cual identifica tiempos del evento en el intervalo $[0, \tau)$. Los eventos que ocurran después del tiempo τ se consideran censurados a la derecha.

Los autores proveen un proceso gausiano bivariado, $W_i(t) = \{W_{1i}(t), W_{2i}(t)\}$, de manera que el modelo de medidas repetidas dependa de $W_{1i}(t)$ y el modelo de tiempo para el evento dependa de $W_{2i}(t)$. El marco de trabajo para el modelamiento incluye métodos estándar de selección de modelos por separado. El análisis univariado del proceso longitudinal y del proceso de tiempo para el evento, respectivamente, es como sigue:

1. El proceso $y_{i1}, y_{i2}, \ldots$ en los tiempos $t_{i1}, t_{i2}, \ldots$ es expresado como

$$Y_{ij} = \mu_i(t_{ij}) + W_{1i}(t_{ij}) + Z_{ij}, \tag{12.3.14}$$

   donde la respuesta media $\mu_i(t_{ij})$ es descrita mediante un modelo lineal, $\mu_i(t) = x'_{1i}\beta_1$; Z_{ij} es una medida de error, la cual se distribuye normal con media cero y varianza σ_Z^2, y $W_1(t)$ es un proceso gausiano que puede ser descompuesto en un término de efectos aleatorios y un término estacionario, es decir,

$$W_{1i}(t) = d_{1i}(t)'U_{1i} + V_{1i}(t). \tag{12.3.15}$$

2. El proceso de intensidad del evento en el tiempo t está dado por el modelo semiparametrico multiplicativo

$$\lambda_i(t) = \lambda_0(t)\exp\{\alpha(t) + W_{2i}(t)\}, \tag{12.3.16}$$

donde $\lambda_0(t)$ es una línea base no-paramétrica, $\alpha(t)$ es un modelo lineal que describe los efectos proporcionales de las covariables medidas en el tiempo t, y el proceso gausiano $W_{2i}(t)$ tiene una descomposición comparable a $W_{1i}(t)$, así:

$$W_{2i}(t) = d_{2i}(t)'U_{2i} + V_{2i}(t). \tag{12.3.17}$$

La posible asociación entre el proceso longitudinal y el tiempo para el evento se modela mediante la postulación de una distribución conjunta para estos dos. La distribución gausiana multivariada para los dos efectos aleatorios, U_1 y U_2, cumple tal papel, con una estructura de covarianzas no diagonal $\gamma_{12}(u) = Cov\{W_1(t), W_2(t-u)\}$ para el proceso bivariado $W(t)$.

Y y N denotan los datos asociados con la medición longitudinal y el proceso de conteo ligado con el tiempo para el evento, respectivamente. Y $\mathcal{W}$ denota la trayectoria completa de $W_1(t)$ para $0 \leq t \leq \tau$. La contribución de un individuo o una unidad a la verosimilitud puede expresarse en la forma

$$l(\theta) = l_1(\theta; Y)\mathcal{E}[l_2(\theta; N|\mathcal{W}_2], \tag{12.3.18}$$

donde $l_1(\theta; Y)$ es la forma estándar correspondiente a la distribución gausiana multivariada de Y. Si los procesos $W_1(;)$ y $W_2(;)$ son independientes, el segundo término se reduce a la forma estándar de un modelo de riesgos proporcionales con término de fragilidad.

Henderson, Diggle y Dobson [31] extienden el método de Wulfsohn y Tsiatis [73] para una amplia clase de modelos y advierten sobre los problemas de identificación, los cuales pueden deberse a que $W_2(t)$ se asume dinámico con una especificación no paramétrica de la intensidad de línea base $\lambda_0(t)$.

Además de los esquemas de modelamiento conjunto para datos longitudinales y tiempo para un evento resumidos anteriormente, hay propuestas bayesianas que emplean la metodología MCMC para ajustar modelos conjuntos a datos que contienen medidas repetidas, tiempo para un evento y covariables [25, 6]. Una revisión del modelamiento conjunto de estos datos es desarrollado por Tsiatis y Davidian [67].

En el libro de Rizopoulos [61] se proporciona una visión general de la teoría y la aplicación de modelos conjuntos para datos longitudinales y de sobrevida. El autor presenta varias extensiones, incluyendo, entre otros, diferentes tipos de estructura de asociación entre datos longitudinales y de sobrevida. Para esto, presenta modelos en los cuales se incluyen factores de estratificación, se incorporan efectos rezagados, se consideran covariables exógenas dependientes del tiempo y se trata con escenarios de riesgos

competentes. Para la parte longitudinal, se adopta un enfoque usual de modelamiento de datos longitudinales continuos y uso de modelos lineales de efectos mixtos. Todos los análisis incluidos en este libro se implementan en el paquete R, con el módulo JM (por su sigla en inglés, *Joint Models*), en un entorno de cómputo estadístico y gráfico. Este paquete se adapta a una amplia gama de modelos conjuntos, incluidas las extensiones mencionadas anteriormente.

El libro de Faucett y Thomas [24] presenta una revisión de la metodología estadística desarrollada en los últimos años para el modelamiento de datos longitudinales y de sobrevida. Además, presenta una discusión sistemática de los modelos conjuntos para datos longitudinales y de tiempo hasta un evento con aplicaciones en varios escenarios.

Específicamente, ese libro proporciona una visión general de varias áreas principales en las que los modelos conjuntos se han desarrollado para abordar importantes preguntas científicas y problemas que no pueden ser respondidos por análisis separados de los datos longitudinales y datos de supervivencia, respectivamente. También incluye el análisis de datos longitudinales de datos faltantes monotónicos atribuidos a eventos continuos o discretos, datos longitudinales con valores perdidos tanto intermitentes como monotónicos, modelos de tiempo al evento con covariables dependientes del tiempo medidas intermitentemente, estudios longitudinales con tiempos de observación informativa y predicción dinámica en modelos conjuntos.

Además, el libro presenta algunas extensiones de los modelos conjuntos para el caso de los tiempos a eventos de riesgos competentes. También aborda los datos longitudinales multivariados conjuntamente modelados con datos de sobrevida multivariados. Incluye temas como el análisis de sensibilidad para investigar el impacto de los supuestos del modelo sobre la inferencia estadística de modelos conjuntos, el diagnóstico de los modelos y la selección de covariables en modelos conjuntos. Finalmente, trata los modelos multiestado conjuntos y los modelos tasa de cura.

Capítulo

trece

Datos de sobrevida multivariados

$$f(t) = \frac{\partial^p F(t)}{\partial t_1 \cdots \partial t_p} = (-1)^p \frac{\partial^p S(t)}{\partial t_1 \cdots \partial t_p}$$

13.1. Introducción

Hasta el capítulo anterior, se han presentado datos relacionados con el tiempo para la ocurrencia de algún evento, por una única vez, sobre una unidad. En el caso estándar o clásico, el evento es la muerte y no sobra decir que esta ocurre solo una vez. Sin embargo, hay otro tipo de eventos que se presentan sobre una unidad más de una vez, por ejemplo, eventos como la enfermedad, el daño en una máquina, los matrimonios, el empleo (o desempleo) y la recesión económica. Estos son eventos que pueden no ocurrir u ocurrir más de una vez sobre la misma unidad. Este tipo de datos se conoce como datos de *sobrevida multivariantes*, lo que significa que son casos en los que más de un tiempo puede ser observado sobre cada unidad.

A continuación, se presentan algunos ejemplos de datos con estructura de datos de tiempo hasta un evento multivariado, orientados principalmente por los textos de Hougaard [33], Crowder [17] y Cook y Lawless [13].

Ejemplo 13.1.1. Un primer caso de datos multivariados corresponde a cuatro grupos, cada uno de 10 ratas. Las ratas del grupo 1 reciben la dosis más baja de un analgésico y las del grupo 4 reciben la dosis más alta. Las ratas fueron expuestas a un estímulo sensorial inofensivo a los 0, 15, 30 y 60 minutos después de la administración del medicamento (origen de cada grupo), y sus tiempos de respuesta en segundos (t_1, t_2, t_3, t_4) se registraron en cada ocasión. Los grupos aparecen como cuatro bloques en la mesa. Hay 160 observaciones, consignadas en la tabla 13.1, de las cuales 17 están censuradas por la derecha a los 10 segundos.

Ejemplo 13.1.2. Los datos surgen de investigaciones sobre el efecto en la función cerebral del plomo absorbido corporalmente. Los datos de la tabla 13.2 surgen de estos estudios. Hay tres grupos de ratas, con diferentes niveles de plomo en el cuerpo y cinco niveles de dosis de un analgésico droga en cada grupo. Los tiempos se registran en segundos hasta que la rata responda a un estímulo táctil inofensivo, cuya intensidad aumenta constantemente desde cero. Se mide dos veces por rata, una antes de la administración del medicamento y una después (t_1 y t_2). Los tiempos se dan con una aproximación de 0.5 segundos y son censurados a derecha en un tiempo de 250 segundos.

Tabla 13.1. Tiempos al evento repetidos

Grupo 1				Grupo 2				Grupo 3				Grupo 4			
t_1	t_2	t_3	t_4	t_1	t_2	t_3	t_4	t_1	t_2	t_3	t_4	t_1	t_2	t_3	t_4
1.58	1.78	2.02	2.19	2.02	3.86	2.73	2.88	2.74	5.10	7.00	3.54	1.68	5.33	10.0	10.0
1.55	1.40	2.20	1.73	1.75	3.38	3.74	2.57	1.89	8.00	9.80	6.63	2.80	5.93	8.77	4.62
1.47	2.66	3.05	3.76	1.93	3.62	6.05	2.91	1.47	9.77	9.98	10.0	2.19	8.73	10.0	8.20
2.16	2.28	2.04	2.12	2.04	4.70	3.70	3.05	3.13	9.04	9.92	6.25	1.52	10.0	10.0	10.0
2.19	2.24	1.99	1.58	2.00	3.34	4.14	2.78	1.53	4.72	5.56	3.25	1.85	7.71	8.35	7.57
2.07	1.97	1.50	1.24	1.63	3.37	4.84	2.36	2.12	4.73	7.90	6.42	2.78	7.41	10.0	10.0
1.28	1.67	2.25	1.54	1.97	3.72	7.83	3.14	1.28	4.32	6.24	4.73	1.81	7.56	10.0	10.0
1.53	2.68	1.79	2.03	2.42	4.81	4.90	1.69	1.50	8.91	9.22	4.30	1.80	10.0	9.41	10.0
2.62	2.15	1.60	2.91	1.38	3.88	4.20	2.05	2.05	4.21	10.0	3.03	1.16	9.95	10.0	10.0
1.67	2.52	1.53	1.98	2.32	3.75	3.06	3.81	1.53	7.10	9.88	3.72	1.67	7.82	10.0	7.97

Ejemplo 13.1.3. Se midieron la longitud y la resistencia a la ruptura de fibras. Las fibras se cortaron en segmentos de 5, 12, 30 y 75 mm de longitud, y cada segmento fue sometido a una carga constantemente creciente hasta su rotura. La carga máxima es 4.0; si la fibra todavía está intacta, en ese punto se libera sin alterar sus características; estas se consideran censuradas a la derecha por 4.0. La tabla 13.3 contiene estos datos.

Ejemplo 13.1.4. El uso racional (ahorro) de la energía eléctrica en el hogar es una de las medidas que deben impulsar los gobiernos, con el fin de obtener credenciales verdes. Para tal fin se instalaron medidores en una serie de hogares para observar los tiempos durante los cuales varios aparatos fueron encendidos. Los tiempos fueron grabados en una fase inicial de cuatro semanas, y luego, durante una fase experimental de cuatro semanas. Durante la última fase, se instalaron dispositivos que apagan los aparatos automáticamente después de un tiempo establecido, con el propósito de evitar el desperdicio de energía cuando un electrodoméstico no está en uso activo. Los datos se muestran en la tabla 13.3 y corresponden a la última semana en cada fase. Los datos son del tipo *sobrevida* en un sentido amplio: los tiempos registran durante cuánto tiempo sobrevive el dispositivo hasta que se apague, y son censurados en 1440, el número de minutos en un día. La ubicación de la lámpara, en la sala, la sala de estar o donde sea, no es registrada en los datos, aunque las diferentes ubicaciones podrían ayudar a explicar la gran variación en el uso entre los hogares. Por otra parte, el es-

tudio se desarrolló en el invierno, y en algunos hogares la lámpara parece dejarse encendida la mayor parte del tiempo (por ejemplo, en el hogar 9).

Tabla 13.3. Tiempos al evento repetidos

	Carga					Carga			
Fibra	5 mm	12 mm	30 mm	75 mm	Fibra	5 mm	12 mm	30 mm	75 mm
1	3.30	3.32	2.39	2.08	11	4.00	4.00	2.99	2.30
2	4.00	3.67	2.49	2.06	12	4.00	3.03	2.80	2.30
3	4.00	4.00	3.16	2.05	13	3.01	3.17	2.41	2.07
4	3.64	2.41	2.20	1.80	14	4.00	2.91	2.18	1.83
5	2.73	2.24	1.91	1.68	15	4.00	3.87	2.24	2.09
6	4.00	4.00	2.74	2.22	16	4.00	3.82	2.59	2.48
7	3.29	3.08	2.44	2.37	17	4.00	4.00	2.40	2.22
8	3.55	2.35	2.38	2.37	18	3.48	2.14	2.35	2.05
9	3.03	2.26	1.64	2.03	19	3.05	2.96	1.91	2.20
10	4.00	4.00	2.98	2.39	20	3.60	2.92	2.42	2.09

Ejemplo 13.1.3. Se midieron la longitud y la resistencia a la ruptura de fibras. Las fibras se cortaron en segmentos de 5, 12, 30 y 75 mm de longitud, y cada segmento fue sometido a una carga constantemente creciente hasta su rotura. La carga máxima es 4.0; si la fibra todavía está intacta, en ese punto se libera sin alterar sus características; estas se consideran censuradas a la derecha por 4.0. La tabla 13.3 contiene estos datos.

Ejemplo 13.1.4. El uso racional (ahorro) de la energía eléctrica en el hogar es una de las medidas que deben impulsar los gobiernos, con el fin de obtener credenciales verdes. Para tal fin se instalaron medidores en una serie de hogares para observar los tiempos durante los cuales varios aparatos fueron encendidos. Los tiempos fueron grabados en una fase inicial de cuatro semanas, y luego, durante una fase experimental de cuatro semanas. Durante la última fase, se instalaron dispositivos que apagan los aparatos automáticamente después de un tiempo establecido, con el propósito de evitar el desperdicio de energía cuando un electrodoméstico no está en uso activo. Los datos se muestran en la tabla 13.3 y corresponden a la última semana en cada fase. Los datos son del tipo *sobrevida* en un sentido amplio: los tiempos registran durante cuánto tiempo sobrevive el dispositivo hasta que se apague, y son censurados en 1440, el número de minutos en un día. La

Tabla 13.2. Tiempos en pares de respuesta: t_1: pre-droga y t_2: post-droga

Dosis 0.0		Dosis 0.1		Dosis 0.2		Dosis 0.4		Dosis 0.8	
Grupo 1									
t_1	t_2	t_1	t_2	t_1	t_2	t_1	t_2	t_1	t_2
47	29	43.5	107	40.5	103.5	45	241	48.5	250
36	35.5	48.5	109	56	203	40	218.5	60	250
51.5	48.5	37	89	57	127.5	52.5	250	39.5	250
39.5	77	55.5	93.5	58.5	114	57	211	39	245.5
31	32.5	27.5	62.5	33.5	95	40.5	117.5	28	224.5
32	49.5	41	71.5	27.5	106	26.5	250	36.5	215.5
35	47.5	36.5	92	40.5	130.5	58.5	179.5	35	249.5
50.5	55	35	79.5	38	131	35.5	193.5	43.5	250
		27.5	74	34.5	115	33	141.5		
		35	98						
Grupo 2									
t_1	t_2	t_1	t_2	t_1	t_2	t_1	t_2	t_1	t_2
46	65	46	72	45.5	83	36	131	35.5	186
38	35.5	53	91	45.5	119	53	107	59	250
54	21.5	40.5	130.5	68.5	121.5	51	130.5	38.5	250
42.5	63	60.5	78	48.5	110	37	144	34.5	250
60.5	40.5	47	101.5	34	83.5	67.5	130	34.5	184.5
38.5	39	47	108	41	138	51	250	56.5	249
41	36.5	36	67	41.5	77	45.5	144	37.5	229
50	45.5	31.5	67.5			49.5	152	33.5	250
Grupo 3									
t_1	t_2	t_1	t_2	t_1	t_2	t_1	t_2	t_1	t_2
70	59.5	67.5	103	66	98	49	126.5	68.5	250
51.5	42.5	50	98.5	42.5	103.5	57	111.5	29	250
38.5	40.5	58	68.5	55.5	85.5	62	99.5	41	204
55.5	64	46.5	70	52	93.5	47.5	145	40	148
59.5	79	59	70	69.5	109	39.5	170.5	65	250
60.5	64.5	71.5	53.5	44.5	114	44	157	62.5	250
29.5	40.5	30	99.5	45.5	89	27.5	120	35	214
42.5	41	29	69	29.5	87	75	150	45.5	197.5
30	35.5	25	70	29.5	93				

Tabla 13.4. Tiempos de uso de energía eléctrica para una lámpara

Hogar	Día (Semana 1)							Día (Semana 2)						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
1	196	165	109	75	37	12	101	158	161	43	90	145	39	75
2	34	44	7	33	38	34	42	5	14	13	13	6	4	3
3	40	59	25	70	45	80	54	48	4	47	3	22	11	6
4	344	61	299	112	22	188	128	87	61	73	21	151	128	69
5	106	49	57	37	103	90	57	22	40	54	87	73	46	25
6	95	82	174	101	104	69	125	53	4	38	75	21	54	41
7	168	137	116	107	62	111	105	48	85	44	58	71	149	49
8	169	163	148	115	154	156	75	24	59	54	74	68	31	81
9	1435	1439	1435	1438	635	45	1136	61	71	52	62	77	45	56
10	80	63	193	91	175	21	42	68	38	60	46	132	75	60
11	68	103	124	102	50	222	11	166	55	39	89	44	47	75
12	188	188	162	155	165	148	86	48	44	75	141	49	87	64
13	17	95	92	107	44	37	97	52	21	36	67	17	40	63
14	70	177	1437	900	1439	487	190	58	79	34	83	94	37	94

ubicación de la lámpara, en la sala, la sala de estar o donde sea, no es registrada en los datos, aunque las diferentes ubicaciones podrían ayudar a explicar la gran variación en el uso entre los hogares. Por otra parte, el estudio se desarrolló en el invierno, y en algunos hogares la lámpara parece dejarse encendida la mayor parte del tiempo (por ejemplo, en el hogar 9).

Ejemplo 13.1.5. Los ataques epilépticos son un ejemplo típico de un evento que puede presentarse una o más veces sobre una misma persona; este tipo de datos se conocen como *datos recurrentes*. Una persona puede experimentar varios, incluso miles de ataques a lo largo de su vida, cada episodio puede tener una duración de unos pocos minutos y no siempre se puede hacer un seguimiento del número total de estas convulsiones. Se presentan datos de este tipo, donde se cuenta el número de eventos dentro de cuatro períodos sucesivos de dos semanas. La tabla 13.5 contiene los datos para los pacientes que recibieron un placebo.

Ejemplo 13.1.6. Gail *et al.* [27] presentaron datos de un experimento de carcinogenicidad (inducción de células cancerígenas) en el que se cuenta el tiempo hasta el desarrollo de tumores mamarios para 48 ratas hembras. Las ratas fueron expuestas a un carcinógeno y condicionadas durante 60 días antes de recibir aleatoriamente un tratamiento o un control. Hubo un

Tabla 13.5. Conteo de episodios de epilepsia. Grupo placebo

Edad	Base	Periodo 1	Periodo 2	Periodo 3	Periodo 4
31	11	5	3	3	3
30	11	3	5	3	3
25	6	2	4	0	5
36	8	4	4	1	4
22	66	7	18	9	21
29	27	5	2	8	7
31	12	6	4	0	2
42	52	40	20	23	12
37	23	5	6	6	5
28	10	14	13	6	0
36	52	26	12	6	22
24	33	12	6	8	4
23	18	4	4	6	2
36	42	7	9	12	14
26	87	16	24	10	9
26	50	11	0	0	5
28	18	0	0	3	3
31	111	37	29	28	29
32	18	3	5	2	5
21	20	3	0	6	7
29	12	3	4	3	4
21	9	3	4	3	4
32	17	2	3	3	5
25	28	8	12	2	8
30	55	18	24	76	25
40	9	2	1	2	1
19	10	3	1	4	2
22	47	13	15	13	12

Tabla 13.6. Tiempo al tumor (en días) en ratas de laboratorio

Id.	Días detección tumor	Id.	Días detección tumor
1	122	1	3, 42, 59, 61$^{(2)}$, 112, 119
2	-	2	28, 31, 35, 45, 52, 59$^{(2)}$, 77, 85, 107, 112
3	3, 88	3	31, 38, 48, 52, 74, 77, 101$^{(2)}$, 119
4	92	4	11, 114
5	70, 74, 85, 92	5	35, 45, 74$^{(2)}$, 77, 80, 85, 90$^{(2)}$
6	38, 92, 122	6	8($^{(2)}$, 70, 77
7	28, 35, 45, 70, 77, 107	7	17, 35, 52, 77, 101, 114
8	92	8	21, 24, 66, 74, 101$^{(2)}$, 114
9	21	9	8, 17, 38, 42$^{(3)}$
10	11, 24, 66, 74, 92	10	52
11	56, 70	11	28$^{(2)}$, 31, 38, 52, 74$^{(2)}$, 77$^{(2)}$, 80$^{(2)}$, 92$^{(2)}$
12	31	12	17, 119
13	3, 8, 24, 35, 92	13	52
14	45, 92	14	11$^{(2)}$, 14, 17, 52, 56$^{(2)}$, 80$^{(2)}$, 107
15	3, 42, 92	15	17, 35, 66, 90
16	3, 17, 52, 80	16	28, 66, 70$^{(2)}$, 74
17	17, 59, 92, 101, 107	17	3, 14, 24$^{(2)}$, 28, 31, 35, 48, 74, 77, 119
18	45, 52, 85, 101, 122	18	21, 28, 45, 56, 63, 80, 85, 92, 101$^{(2)}$, 119
19	92	19	28, 35, 52, 59, 66$^{(2)}$, 90, 97, 119
20	21, 35	20	8$^{(2)}$, 24, 42, 45, 59, 63$^{(2)}$, 77, 101, 119, 122
21	24, 31, 42, 48, 70, 74	21	80
22	-	22	92, 122$^{(2)}$
23	31	23	21
		24	3, 28, 74
		25	24, 74, 122

período de seguimiento de 122 días, que inició después de la asignación aleatoria del tratamiento, durante el cual se examinaron las ratas cada pocos días para detectar el desarrollo de nuevos tumores. Los datos proporcionados en Gail *et al.* [27] se muestran en la tabla 13.7. Se escriben los tiempos correspondientes a los días en que los tumores fueron detectados.

Los datos de sobrevivencia multivariados, como los anteriores, tienen una característica común y fundamental y es que, por ser mediciones o registros realizados sobre la misma unidad o por ser unidades reunidas o agrupadas en conglomerados, la independencia entre los tiempos de sobrevivencia no se puede asumir, pues estas son dependientes. Esta dependencia es la que debe hacerse visible, considerarse e incluirse en las respectivas es-

timaciones, y, en general, en el modelamiento de estos datos multivariados. Es común emplear el término *tiempos de sobrevivencia correlacionados* para referirse a datos de sobrevida multivariados.

Dentro de la clasificación de los datos multivariados, se consideran dos estructuras principales: *paralelos* y *longitudinales*. La principal diferencia entre estas estructuras de datos es que en los datos paralelos el número de tiempos es fijado por el diseño o planeamiento del estudio, mientras que en los datos longitudinales los puntos en el tiempo se presentan de manera aleatoria.

13.1.1. Datos paralelos

En este caso, varias unidades (sujetos, unidades experimentales o unidades de un sistema) son seguidas en forma simultánea. Esto produce un conjunto de datos en forma matricial. Estas respuestas tienen la forma T_{ij}, con $i = 1, \ldots, n$ siendo el número de unidades y $j = 1, \ldots, k$, la ordinalidad de los tiempos donde se registra el evento. Se asume que hay independencia entre diferentes unidades, en cambio, posiblemente, hay dependencia entre los tiempos para la misma unidad i. Aunque el número de tiempos, k, sobre una unidad es determinado previamente, algunos de estos tiempos corresponden a tiempo al evento y otros a censuras, en consecuencia, el número de eventos no es especificado; naturalmente, el número de eventos es, a lo más, k.

Un ejemplo de estos datos corresponde a la medición del tiempo (en semanas) hasta la aparición de un tumor en ratas sometidas a diferentes tipos de drogas [33, p. 14].

13.1.2. Datos longitudinales

Se asume que se observa un proceso estocástico, uno para cada valor de i, y se registran los eventos o transiciones en el tiempo. En el caso de las personas, habrá un proceso por cada una; cada uno de los cuales se denomina historia de vida. Se asume la independencia entre los i. Así, los tiempos de transición serán una sucesión creciente de tiempos, y en el caso más simple, únicamente el último tiempo estará sujeto a censura, todos los demás son tiempos de transición. El número de tiempos, K_i, es aleatorio. Los tiempos pueden corresponder a la recurrencia de eventos similares, estos se denominan *eventos recurrentes*, o pueden ser transiciones entre varios estados, en cuyo caso, los datos incluyen no solo el tiempo, sino el nombre del estado a donde el proceso llega.

Tabla 13.7. Tiempo al tumor (en días) en ratas de laboratorio

Droga	Control l	Control 2	Droga	Control l	Control 2
101+	49	104+	104+	102+	104+
104+	104+	104+	77+	97+	79+
89+	104+	104+	88	96	104+
104	94+	77	96	104+	104+
82+	77+	104+	70	104+	77+
89	91+	90+	91+	70+	92+
39	45+	50	103	69+	91+
93+	104+	103+	85+	72+	104+
104+	63+	104+	104+	104+	74+
81+	104+	69+	67	104+	68
104+	104+	104+	104+	104+	104+
104+	83+	40	87+	104+	104+
104+	104+	104+	89+	104+	104+
78+	104+	104+	104+	81	64
86	55	94+	34	104+	54
76+	87+	74+	103	73	84
102	104+	80+	80	104+	73+
45	79+	104+	94	104+	104+
104+	104+	104+	104+	101	94+
76+	84	78	80	81	76+
72	95+	104+	73	104+	66
92	104+	102	104+	98+	73+
55+	104+	104+	49+	83+	77+
89	104+	104+	88+	79+	99+
103	91+	104+	104+	104+	79

Cuando se tiene que una unidad puede experimentar el mismo evento varias veces, se está tratando con *eventos recurrentes*. Este caso incluye, por ejemplo, eventos de hipoglucemia para un paciente diabético. Puede haber un solo tipo de evento, como en el caso de la fertilidad (nacimientos vivos), o el proceso puede alternar entre dos estados, por ejemplo, en el ciclo menstrual pueden considerase cada uno: el sangrado o no sangrado. El número de eventos por periodo de tiempo de cada individuo puede ser bajo, por ejemplo, los nacimientos, o puede ser alto, como en el caso de ataques de epilepsia.

13.2. Distribuciones de sobrevida multivariadas

Sea $\mathbf{T} = (T_1, T_2, \ldots, T_p)$ un vector de tiempos a un evento, también se puede considerar como un conjunto de variables aleatorias no negativas. La *función de distribución conjunta* es definida como

$$F(\mathbf{t}) = P(\mathbf{T} \leq \mathbf{t}) = P(T_1 \leq t_1, T_2 \leq t_2, \cdots, T_p \leq t_p), \tag{13.2.1}$$

donde $\mathbf{t} = (t_1, \ldots, t_p)$.

Similarmente, *la función de sobrevida conjunta* es

$$S(\mathbf{t}) = P(\mathbf{T}t@ > \mathbf{t}) = \mathbf{1} - \mathbf{F(t)} = \mathbf{P(T_1 > t_1, T_2 > t_2, \cdots, T_p > t_p)}. \tag{13.2.2}$$

Si los T_j son, conjuntamente, absolutamente continuos, entonces la función de densidad conjunta, $f(t) = f(t_1, \ldots, t_p)$, es

$$f(t) = \frac{\partial^p F(\mathbf{t})}{\partial t_1 \cdots \partial t_p} = (-1)^p \frac{\partial^p S(\mathbf{t})}{\partial t_1 \cdots \partial t_p}. \tag{13.2.3}$$

Para subconjuntos de los componentes del vector aleatorio $\mathbf{T}$, las funciones de distribuciones marginal son dadas por

$$F_j(t) = P(T_j \leq t), \; F_{jk}(t_j, t_k) = P(T_j \leq t_j, T_k \leq t_k), \quad \text{y así sucesivamente.} \tag{13.2.4}$$

De manera similar, las funciones de sobrevida asociadas con subconjuntos de componentes del vector aleatorio $\mathbf{T}$ son

$$S_j(t) = P(T_j > t), \; S_{jk}(t_j, t_k) = P(T_j > t_j, T_k > t_k), \quad \text{y así sucesivamente.} \tag{13.2.5}$$

Considere que el vector $\mathbf{T}$ se particiona como $\mathbf{T} = (\mathbf{T}_{(1)}, \mathbf{T}_{(2)})$. Entonces,

$$P(\mathbf{T}_{(2)}t@ > \mathbf{t_{(2)}}|\mathbf{T}_{(1)}\mathbf{t}@ > \mathbf{t}_{(1)}) = \frac{\mathbf{S(t)}}{\mathbf{S(t_{(1)})}}. \tag{13.2.6}$$

Las variables aleatorias T_j, contenidas en $\mathbf{T}$, son *independientes* si

$$F(\mathbf{t}) = F(t_1, t_2, \ldots, t_p) = F_1(t_1)F_2(t_2)\cdots F(t_p) = \prod_{j=1}^{p} F_j(t_j) \text{ equivale a} \tag{13.2.7}$$

$$F(\mathbf{t}) = S(t_1, t_2, \ldots, t_p) = S_1(t_1)S_2(t_2)\cdots S(t_p) = \prod_{j=1}^{p} S_j(t_j). \tag{13.2.8}$$

Cuando las variables aleatorias T_j no son independientes, existen varias formas de medir esta dependencia, una de las cuales es mediante la probabilidad condicional mostrada en (13.2.6); el problema en este caso es decidir sobre los componentes de $\mathbf{T}$ que conformarían a $\mathbf{T}_{(1)}$ y $\mathbf{T}_{(2)}$, respectivamente. Dos medidas alternativas del grado de dependencia son las razones

$$\mathscr{D}_1 = \frac{F(\mathbf{t})}{\prod_{j=1}^{p} F_j(t_j)} \text{ y } \mathscr{D}_2 = \frac{S(\mathbf{t})}{\prod_{j=1}^{p} S_j(t_j)}, \tag{13.2.9}$$

si $\mathscr{D}_1 > 1$ (o $\mathscr{D}_2 > 1$) para todo $\mathbf{t}$ (o $\mathscr{D}_2 < 1$) hay dependencia positiva, y si $\mathscr{D}_1 < 1$ hay dependencia negativa.

La *función de riesgo* en un componente de $\mathbf{T}$ es

$$h_j(t_j; \mathbf{t}) = \lim_{\Delta \to 0} \frac{P(T_j \leq t_j + \Delta | \mathbf{T}t@ > \mathbf{t})}{\Delta}. \tag{13.2.10}$$

Por lo tanto, se puede observar a un grupo de unidades y tener un interés particular en el la j-ésima unidad.

La función escalar de riesgo multivariado se define como

$$h(\mathbf{t}) = \frac{f(\mathbf{t})}{S(\mathbf{t})}. \tag{13.2.11}$$

En el caso bivariado ($p = 2$), la función de riesgo es

$$h(t_1, t_2) = \frac{f(t_1, t_2)}{S(t_1, t_2)} = \frac{\partial^2 S(t_1, t_2)}{\partial t_1 \partial t_2} / S(t_1, t_2). \tag{13.2.12}$$

El miembro derecho de la igualdad (13.2.12) no es la segunda derivada del logaritmo de la función de sobrevida bivariada ($\ln S(t_1, t_2)$), en consecuencia, esta expresión no es una extensión del caso univariado; esto se puede confrontar con (1.5.3).

13.3. Distribución exponencial bivariada de Gumbel

Una de las tres formas de la distribución exponencial bivariada de Gumbel, propuestas por Gumbel [29], es

$$f(\mathbf{t}) = \exp\{\theta_1 t_1 - \theta_2 t_2 - \nu_1 t_1 t_2\}, \tag{13.3.1}$$

donde $\theta_1 > 0$, $\theta_2 > 0$, el *parámetro de dependencia*, ν, debe satisfacer $0 < \nu\theta_1\theta_2$, y un valor de $\nu = 0$ refleja la independencia entre T_1 y T_2. La razón de dependencia $\mathcal{D}_1$ es

$$\mathcal{D}_1 = \frac{\exp\{-\theta_1 t_1 - \theta_2 t_2 - \nu_1 t_1 t_2\}}{\exp\{-\theta_1 t_1\}\exp\{-\theta_2 t_2\}} = \exp\{-\nu t_1 t_2\}. \tag{13.3.2}$$

Como $\exp\{-\nu t_1 t_2\}$ es siempre menor que 1, entonces la distribución muestra una dependencia negativa.

La función de sobrevida marginal $S_j(t_j) = \exp\{-\theta_j t_j\}$, para $j = 1, 2$, y la condicional de T_2 dado T_1, es

$$P(T_2 > t_2 | T_1 > t_1) = \exp\{-\theta_2 t_2 - \nu t_1 t_2\ \}$$

y

$$P(T_2 > t_2 | T_1 = t_1) = (1 + \nu t_2/\theta_1)\exp\{-\theta_2 t_2 - \nu t_1 t_2\ \}.$$

Estas probabilidades son decrecientes en t_1, lo cual refleja una dependencia negativa, y para el caso donde se conoce que $T_1 = t_1$, en lugar de $T_1 > t_1$ se incrementa la probabilidad de que $T_2 > t_2$ por un factor $(1 + \nu t_1/\theta_1) > 1$.

13.4. Modelos multivariados de sobrevida

13.4.1. Distribución log-normal multivariada

La distribución normal multivariada tiene asociada la matriz de correlación como uno de los dos parámetros que la identifican; esta matriz puede asumir diferentes estructuras de correlación. Su adaptación para el análisis de sobrevida se logra mediante la transformación exponencial; esta transformación produce la distribución log-normal.

Sea $\mathbf{Y}$ con distribución normal p-variante, es decir, $\mathbf{Y} \sim N_p(\mu, \Sigma)$, donde μ es el vector de medias y Σ es la matriz de covarianzas; en la diagonal de Σ están las varianzas σ_j^2 y fuera de la diagonal, las covarianzas σ_{jk}, como

se puede consultar en Díaz y Morales [21, cap. 2]. Entonces $\mathbf{T}$, definido componente a componente por $T_j = \exp(Y_j)$, tiene una distribución *log-normal p-variante*. La media de los componentes son de la forma $E(T_j) = e^{\mu_j + \sigma_j^2/2}$ y la varianza $var(T_j) = (e^{\sigma_j^2} - 1)\mu_j^2$.

En algunos estudios del efecto de algún tratamiento, las observaciones son frecuentemente del tipo "pre-post" tratamiento. Por ejemplo, la respuesta a un medicamento antes y después de su aplicación. En estos casos, la atención se dirige a una distribución bivariada.

Suponga que $\mathbf{T} = (T_1, T_2)'$ tiene distribución $N_2(\mu, \Sigma)$, con media $\mu = (\mu_1, \mu_2)'$ y matriz de covarianza

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix},$$

donde $\rho = \sigma_{12}/(\sigma_1\sigma_2)$ es el coeficiente de correlación. La función de densidad normal bivariada de $\mathbf{T}$ es

$$\phi_2(y; \mu, \Sigma) = |2\pi\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right\},$$

y la distribución para las variables estandarizadas mediante $Z_j = (Y_j - \mu_j)/\sigma_j$ es

$$\phi_2(z; 0, R) = [2\pi\sqrt{(1-\rho^2)}]^{-1} \exp\left\{-\frac{1}{2}(z_1^2 - 2\rho z_1 z_2 + z_2^2)/(1-\rho^2)\right\},$$

donde $R = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ es la matriz de correlación.

13.4.2. Distribución exponencial bivariada en la versión de Freund (1961)

En un sistema de dos componentes, los tiempos de vida, T_1 y T_2, son en principio variables aleatorias exponenciales independientes. No obstante, la falla de un componente puede alterar el desempeño futuro del otro, de manera que se debe cambiar el parámetro exponencial θ_j a μ_j $(j = 1, 2)$. El caso $\mu_j > \theta_j$ sería apropiado cuando la sobrecarga en el componente que no falló ("sobreviviente") aumenta después de la falla, pues esta carga, antes de la falla, era compartida por cada sistema. Recíprocamente, si $\mu_j < \theta_j$, reflejaría cierto alivio del estrés o carga, en comparación a cuando los componentes anteriormente compartían el esfuerzo o carga de trabajo.

La función de sobrevida conjunta se puede expresar como

$$S(\mathbf{t}) = (\theta_+ - \mu_2)^{-1} \left\{\theta_1 e^{-(\theta_+ - \mu_2)t_1 - \mu_2 t_2} + (\theta_2 - \mu_2)e^{-\theta_+ t_2}\right\}, \qquad (13.4.1)$$

para $t_1 \leq t_2$, condicionado a que $\theta_+ = \theta_1 + \theta_2 \neq \mu_2$. La forma para $t_1 \geq t_2$ se obtiene desde (13.4.1) intercambiando θ_1 y θ_2, μ_1 y μ_2, t_1 y t_2. Considerando $t_1 = 0$ en la forma de $S(\mathbf{t})$, se obtiene la marginal

$$S(\mathbf{t}) = (\theta_+ - \mu_2)^{-1} \left\{ \theta_1 e^{-\mu_2 t_2} + (\theta_2 - \mu_2) e^{-\theta_+ t_2} \right\}. \tag{13.4.2}$$

La función de sobrevida (13.4.2) es una mezcla de dos exponenciales, por lo tanto, la distribución conjunta no es una distribución exponencial bivariada. La independencia entre T_1 y T_2 se obtiene si y solo si $\mu_1 = \theta_1$ y $\mu_2 = \theta_2$.

13.4.3. Distribución exponencial bivariada en la versión de Marshall-Olkin (1967)

Considere un sistema con dos componentes expuesto a tres tipos de eventos que dejan sin funcionamiento a los componentes. El primer tipo de evento ocurre sobre el componente 1, el segundo evento se presenta sobre el componente 2 y el tercero sobre los dos componentes. Suponga que los eventos se presentan de acuerdo con procesos de Poisson con tasas de ocurrencia λ_1, λ_2 y λ_{12}, respectivamente. Sea $N_j(t)$ el número de eventos del tipo j que ocurren durante el periodo de tiempo $(0, t)$. Entonces,

$$\begin{aligned} S(\mathbf{t}) &= P(T_1 > t_1, T_2 > t_2) = P[N_1(t_1) = 0, N_2(t_2) = 0, N_3\{\text{máx}(t_1, t_2)\} = 0] \\ &= e^{-\lambda_1 t_1} e^{-\lambda_2 t_2} e^{-\lambda_{12} \text{máx}\{t_1, t_2\}}. \end{aligned} \tag{13.4.3}$$

13.5. Fragilidad

En la sección 7.5 se trata el modelo de Cox extendido a la forma y estructura de un modelo de fragilidad. Para desarrollar los modelos de fragilidad multivariados se inicia con la función de sobrevida marginal de T_j. De acuerdo con (1.5.4), es $S_j(t) = \exp -H_j(t)$, donde $H_j(t)$ es la función de riesgo acumulada de T_j. Suponga que

a) La H_j (para $j = 1, \ldots, p$) es una unidad que comparte un factor común aleatorio z, y que

b) z varía en una población de unidades con distribución K en $(0, \infty)$.

En algunos contextos, z se conoce como la *fragilidad* de la unidad. Ahora se reemplaza $H_j(t)$ por $zH_j(t)$ y se hace el supuesto de que los T_j para una unidad sean condicionalmente independientes dado z. La función de sobrevida

conjunta, dado z, es

$$S(\mathbf{t}|\mathbf{z}) = \exp\left\{-\sum_{j=1}^{p} H_j(t)\right\}. \tag{13.5.1}$$

La función de sobrevida conjunta no condicional se calcula desde la integral sobre la distribución de z como

$$S(\mathbf{t}) = \int_0^\infty e^{-zs} dK(z) = \mathscr{L}_K(s), \tag{13.5.2}$$

donde $s = H_1(t_1) + \cdots + H_p(t_p)$, $\mathscr{L}_K(\cdot)$ es la transformada de Laplace y la función generadora de momentos de la distribución K. La construcción produce una distribución multivariada en la cual los componentes de tiempo al evento, T_j, en general son dependientes. Sin embargo, la forma de dependencia está restringida por la simetría de S en la H_j.

En el caso bivariado, su construcción, basada en la función de sobrevida, es

$$S(\mathbf{t}|\mathbf{z}) = \exp\{-z_1 H_1(t_1) - z_2 H_2(t_2)\} \tag{13.5.3}$$

y

$$S(\mathbf{t}) = \int_0^\infty e^{-z_1 H_1(t_1) - z_2 H_2(t_2)} dK(z), \tag{13.5.4}$$

donde $K(\mathbf{z})$ es la función de distribución de $\mathbf{Z} = (Z_1, Z_2)$. La marginal para T_1 se obtiene haciendo $t_2 = 0$, es decir, $S(t_1) = \int e^{-z_1 H_1(t_1)} dK(z_1)$. Para la versión basada sobre las funciones de distribución, se debe reemplazar $e^{-z_j H_j(t_j)}$ por $F_j(t_j)^{z_j}$ $(j = 1, 2)$; esto no produce el mismo resultado que la versión $S(\cdot)$, dado que $S(t_j)^{z_j} \neq 1 - F_j(t_j)^{z_j}$.

13.5.1. Algunos generadores de distribuciones de fragilidad

Se consideran casos particulares al tomar una forma para el riesgo integrado: $H_j(t) = (t/\xi_j)^{\phi_j}$ y, por lo tanto, $s = \sum_j^p (t/\xi_j)^{\phi_j}$. Las siguientes selecciones de K resultan apropiadas.

13.5.1.1. Burr multivariada-BM

Considere que K tiene una distribución gama con parámetro de forma ν y parámetro de escala ζ. A partir de esto se obtiene $S(\mathbf{t}) = (1 + \zeta s)^{-\nu}$, la función de sobrevida de la distribución *BM* (Burr multivariada). Se puede considerar $\zeta = 1$, de modo que el parámetro de escala queda subsumido

en los ξ_j's. La función de sobreviviente marginal del componente T_A en (T_A, T_B) es $S_A(\mathbf{t}_A) = (1+s_A)^{-\nu}$, de la forma BM. Una función sobreviviente condicional sigue como

$$P(\mathbf{T}_B t@ > \mathbf{t_B} | \mathbf{T_A} t@ > \mathbf{t_A}) = (\mathbf{1}+\mathbf{s})^{-\nu} \div (\mathbf{1}+\mathbf{s_A})^{-\nu} = (\mathbf{1}+\mathbf{s}'_\mathbf{B})^{-\nu}, \quad (13.5.5)$$

donde $s'_B = s_B/(1 + s_A)$, que también es de la forma MB, de dimensión $dim(\mathbf{t}_B)$, y con el parámetros ξ modificados conforme a $\xi'_j = \xi_j/(1+s_A)$. La otra función de sobrevida condicional es

$$P(\mathbf{T}_B t@ > \mathbf{t_B} | \mathbf{T_A} = \mathbf{t_A}) = (\mathbf{1} + \mathbf{s}'_\mathbf{B})^{-\nu-\mathbf{a}}, \quad (13.5.6)$$

con $a = dim(\mathbf{t}_A)$. Nuevamente, esta es una distribución de tipo BM.

13.5.1.2. Weibull multivariada

Considerando a K como una distribución estable positiva con exponente característico $\nu \in (0, 1)$, se produce $S(\mathbf{t}) = \exp(-s^\nu)$, que es la función de sobrevida de esta distribución Weibull multivariada. Con la partición $\mathbf{T} = (\mathbf{T}_A, \mathbf{T}_B)$ como en el caso anterior, la función de sobrevida marginal de $\mathbf{T}_A$ es $S_A(\mathbf{t}_A) = \exp(-s_A^\nu)$, también de la forma Weibul multivariada. La primera función de sobrevida condicional es

$$P(\mathbf{T}_B t@ > \mathbf{t_B} | \mathbf{T_A} t@ > \mathbf{t_A}) = \exp(-\mathbf{s}^\nu) \div \exp(-\mathbf{s}_\mathbf{A}^\nu) = \exp(\mathbf{s}_\mathbf{A}^\nu - \mathbf{s}^\nu), \quad (13.5.7)$$

la cual no es Weibull multivariada. No obstante, se extiende la especificación básica mediante la introducción del parámetro $\kappa > 0$, la función de sobrevida ahora es $S(\mathbf{t}) = \exp\{\kappa^\nu - (\kappa + s)^\nu\}$. La función de sobrevida marginal es $S_A(\mathbf{t}_A) = \exp\{\kappa^\nu - (\kappa + s_A)^\nu\}$ y la condicional es

$$\begin{aligned} P(\mathbf{T}_B t@ > \mathbf{t_B} | \mathbf{T_A} t@ > \mathbf{t_A}) &= \exp\{\kappa^\nu - (\kappa + s)^\nu\} \div \exp\{\kappa^\nu - (\kappa + s_A)^\nu\} \\ &= \exp\{(\kappa + s_A)^\nu - (\kappa + s)^\nu\}. \end{aligned} \quad (13.5.8)$$

Las dos funciones de sobrevida, marginal y condicional, son miembros de esta familia de distribuciones extendida. La otra función de sobrevida condicional, $P(\mathbf{T}_B t@ > \mathbf{t_B} | \mathbf{T_A} = \mathbf{t_A})$, no pertenece a esta familia.

13.5.1.3. Beta-geométrica multivariada

Una versión multivariada de la distribución beta-geométrica puede ser construida de la siguiente manera. Suponga que, condicionado al efecto aleatorio z, que suele asociarse con la *fdp* normal, los tiempos al evento discretos $T_1, \ldots, T_p$ son variables geométricas independientes con $P(T_j > t) =$

$(z\rho_j)^t$ para $t = 1, \ldots$. Si z tiene distribución $beta(\phi, \psi)$ a través de las unidades, la función de sobrevida conjunta no condicional del T_j es

$$\begin{aligned} P(\mathbf{T}t@ > \mathbf{t}) &= \int_0^1 \left\{ \prod_{j=1}^{p} (z\rho_j)^{t_j} B(\phi, \psi) z^{\phi-1} (1-z)^{\psi-1} dz \right\} \\ &= \left(\prod_{j=1}^{p} \rho_j{}^{t_j} \right) B(\phi + t_+, \psi)/B(\phi, \psi) \\ &= \left(\prod_{j=1}^{p} \rho_j{}^{t_j} \right) \prod_{l=1}^{t_+} \left(\frac{\phi + l - 1}{\phi + \psi + l - 1} \right), \end{aligned} \tag{13.5.9}$$

donde $t_+ = t_1 + \cdots + t_p$; realmente debería notarse t_{i+} para resaltar que este varía de una unidad a otra. Se denota $a_m = \prod_{l=1}^{m} \left(\frac{\phi+l-1}{\phi+\psi+l-1} \right)$ y con esto se tiene que $a_m = a_{m-1}(\phi+m-1)/(\phi+\psi+m-1)$ inicia en $a_1 = \phi/(\phi+\psi)$. Los a_m se pueden construir recursivamente en un programa para computador y se termina hasta el valor máximo de t_+ de alguna unidad $i = 1, \ldots, n$.

Las funciones de sobrevida marginal son de la forma

$$S_j(t) = \rho_j^t a_t \tag{13.5.10}$$

para $j = 1, \ldots, p$. Una medida de dependencia es $\mathcal{D}(\mathbf{t}) = a_{t_+} / \prod_{j=1}^{p} a_{t_j}$.

13.5.1.4. Gama-Poisson multivariada

La distribución de Poisson no es una distribución de sobrevida natural. Sin embargo, se considera un proceso de Poisson que genera las probabilidades de ocurrencia de un determinado evento, cuya tasa de ocurrencia es λ. Suponga que, en cada oportunidad, el evento puede ocurrir con una probabilidad p y puede no ocurrir con una probabilidad $1 - p$. El tiempo discreto (número de oportunidades) para que ocurra el primer evento tiene una distribución de Poisson con media $(\lambda p)^{-1}$.

Para una versión multivariada, suponga que los tiempos de espera $T_1, \ldots, T_p$ son condicionalmente independientes dado el efecto aleatorio z, con T_j distribuida Poisson con media $(\lambda_j z)^{-1}$. El proceso mezcla una distribución gama y una Poisson. En el caso anterior se mezcla la distribución beta en (0, 1) de la variable aleatoria z, la cual es proporcional a $z^{\alpha}(1 - z)$; integrando sobre z, resulta la distribución de T_j anterior. Ahora suponga que z tiene distribución $gama(\xi, \nu)$. Entonces $T - j$ tiene función de sobrevida

conjunta

$$S(\mathbf{T}=\mathbf{t}) = \int_0^\infty \left\{ \prod_{j=1}^{p} e^{-z\lambda_j}(z\lambda_j)^{t_j}/t_j! \right\} \Gamma(\nu)^{-1}\xi^{-\nu}z^{\nu-1}e^{-z/\xi}dz$$

$$= \left\{ \prod_{j=1}^{p} (\lambda_j)^{t_j}/t_j! \right\} \xi^{t_+}(1+\xi\lambda_+)^{-\nu-t_+}\Gamma(\nu)^{-1}\Gamma(\nu+t_+), \quad (13.5.11)$$

donde $t_+ = t_1 + \cdots + t_p$ y $\lambda_+ = \lambda_1 + \cdots + \lambda_p$.

13.5.1.5. Familia de Marshall-Olkin

Esta familia fue propuesta por Marshall y Olkin [52] mediante la construcción que se resume a continuación. Dadas dos ditribuciones univariadas F_1 y F_2 y una distribución bivariada $G(z_1, z_2)$, se construye

$$H(t_1, t_2) = \int F_1(t_1)^{z_1}F_2(t_2)^{z_2}dG(z_1, z_2). \quad (13.5.12)$$

Las variables z_1 y z_2 se consideran como los efectos aleatorios.

Su inclusión en esta construcción es equivalente a la consideración de las fragilidades anteriores, las cuales se muestran como factores multiplicativos para la función de riesgo que se muestra como alternativa de tipo *Lehmann* de F_1 y F_2, descritas en (5.2.5). Como un caso especial, se puede hacer $z_1 = z_2$, reemplazando $G(z_1, z_2)$ por una función univariada, $G(z)$. Las funciones de sobrevida, S_j, se pueden usar en lugar de las funciones de distribución en la expresión (13.5.12). La extensión a una dimensión mayor se puede hacer desde este caso bivariado.

Ejemplo 13.5.1. Los datos de la tabla 13.8 corresponden al número de copulas de una mujer hasta su primer y segundo embarazo. Para estos datos se asume un modelo probabilístico beta-geométrico. Las frecuencias de los datos de la tabla 13.8 fueron simulados. Así, por ejemplo, en la fila 3 y la columna 4 la frecuencia 8 (en recuadro) corresponde a 8 mujeres a las que les fueron necesarias 3 y 4 cópulas para su primer y segundo embarazo, respectivamente. El valor 7^+ significa que fueron necesarias 7 o más copulas para el embarazo.

Una verificación vía ji-cuadrado evidencia una asociación significativa entre el número de cópulas para el primer y segundo embarazo en la tabla de contingencia 7×7, dado que el valor de la estadística ji-cuadrado de 36 grados de libertad es igual a 54.80, con el correspondiente valor p igual 0.023.

Tabla 13.8. Número de cópulas hasta primer y segundo embarazo

Embarazo 1	Embarazo 2						
	1	2	3	4	5	6	7^+
1	49	21	21	14	10	3	12
2	38	25	10	12	6	1	24
3	16	14	10	8	4	3	11
4	11	7	7	6	6	1	8
5	14	7	2	4	3	1	8
6	6	1	1	6	1	0	3
7^+	18	14	10	3	7	6	27

La función de sobrevida conjunta, de acuerdo con (13.5.9), es

$$S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2) = \left(\rho_1^{t_1} \rho_2^{t_2}\right) a_{t_1+t_2},$$

para $t_1, t_2 = 1, 2, \ldots$, y las funciones de sobrevida marginales $S_j(t) = \rho_j^t a_t$ $(j = 1, 2)$.

Las contribuciones a la función de verosimilitud son como se muestran a continuación. Sea $p(j, k) = P(T_1 = j, T_2 = k)$. Entonces, para $j, k = 1, 2, \ldots$,

$$p(1, j) = 1 - S_1(1) - S_2(j) + S(1, j) - \sum_{k=1}^{j-1} p(1, k),$$

se considera que $\sum_{k=1}^{0}$ es igual a 0; una fórmula similar se tiene para $p(j, 1)$. Para $j, k = 2, 3, \ldots$,

$$p(j, k) = S(j-1, k-1) - S(j-1, k) - S(j, k-1) + S(j, k).$$

La probabilidades marginales a derecha (marginales fila) de la tabla de contingencia, para $j = 2, \ldots, 7$, son

$$p(j, 7^+) = P(T_1 = j, T_2 \geq 7) = S(j-1, 6) - S(j, 6),$$

y las probabilidades marginales inferiores (marginales columna) son

$$p(7^+, j) = P(T_1 \geq 7, T_2 = j) = S(6, j-1) - S(6, j).$$

13.6. Estructura de cópulas

Suponga que (T_1, T_2) tiene la función de distribución $F(t_1, t_2)$, y sean $F_1(t_1)$ y $F_2(t_2)$ las funciones de distribución marginales. Por el teorema de la transformación integral, estas marginales son uniformes en (0, 1). Así, usando $U_j = F_j(T_j)$, para $j = 1, 2$, se tiene

$$\begin{aligned} P(U_1 \le u_1, U_2 \le u_2) &= P[F_1(T_1) \le u_1, F_2(T_2) \le u_2] \\ &= P[T_1 \le F_1^{-1}(u_1), T_2 \le F_2^{-1}(u_2)] \\ &= F[F_1^{-1}(u_1), F_2^{-1}(u_2)]. \end{aligned} \tag{13.6.1}$$

Se ha asumido que las funciones inversas F_1^{-1} y F_2^{-1} están bien definidas, dado que se considera que estas son funciones continuas y estríctamente monótonas crecientes. Así, en la fórmula para $F(t_1, t_2)$, se debe reemplazar t_1 por $F_1^{-1}(u_1)$ y t_2 por $F_2^{-1}(u_2)$.

Lo que se ha hecho aquí es considerar las distribuciones marginales sin alterar lo esencial de la distribución conjunta, es decir, la estructura en conjunto se ha mantenido. Por ejemplo, la medida de dependencia $\mathcal{D}$ no cambia bajo la relación de correspondencia $u_j F_j(t_j)$:

$$\begin{aligned} \mathcal{D}_U(u_1, u_2) &= P(U_1 \le u_1, U_2 \le u_2)/\{P(U - 1 \le u_1)P(U_2 \le u_2)\} \\ &= P(T_1 \le t_1, T_2 \le t_2)/\{P(T_1 \le t_1)P(T_2 \le t_2)\} \\ &= \mathcal{D}_T(t_1, t_2). \end{aligned} \tag{13.6.2}$$

Sin embargo, las medias, las varianzas, las correlaciones, entre otras, generalmente cambiarán por transformaciones no lineales.

Bajo la correspondencia $u_j\ F_j(t_j)$, se puede escribir

$$\begin{aligned} F(t_1, t_2)6 &= P(T_1 \le t_1, T_2 \le t_2) = P(U_1 \le u_1, U_2 \le u_2) \\ &= \mathcal{C}(u_1, u_2) \\ &= \mathcal{C}F_1(t_1), F_2(t_2), \end{aligned} \tag{13.6.3}$$

donde $\mathcal{C}$ es la *función de cópula*. La cópula expresa la función de distribución conjunta en términos de los marginales. Tiene una específica, y por lo tanto restringida, *estructura de dependencia*, pero con marginales opcionales. También se puede basar en el desarrollo de funciones de sobrevida en lugar de funciones de distribución acumulada. Esta idea puede ser extendida a más de dos componentes.

Formalmente, una *cópula* es una distribución multivariada con marginales uniformes estándar (es decir, cada uniforme sobre el intervalo (0, 1)).

Por otra parte, aunque cualquier distribución multivariada puede generar una cópula, mediante una transformación marginal parece haber un requisito implícito de que la función $\mathscr{C}$ debe tener forma simple.

Se considera el cuadrado unitario, $\mathbf{I}^2$, al conjunto definido por los puntos resultado del producto cartesiano del intervalo $\mathbf{I} = [0, 1]$ por sí mismo, es decir, $\mathbf{I}^2 = [0, 1] \times [0, 1]$.

Así, formalmente, la *cópula* es una función de $\mathbf{I}^2$ en $\mathbf{I}$, que tiene las siguientes propiedades:

1. Para todo $u, v \in \mathbf{I}$,
$$\mathscr{C}(u, 0) = 0 = \mathscr{C}(0, v)$$
y
$$\mathscr{C}(u, 1) = u \text{ y } \mathscr{C}(1, v) = v;$$
2. Para todo For u_1, u_2, v_1, v_2 en $\mathbf{I}$, tal que $u_1 \leq u_2$ y $v_1 \leq v_2$,
$$\mathscr{C}(u_2, v_2) - \mathscr{C}(u_2, v_1) - \mathscr{C}(u_1, v_2) + \mathscr{C}(u_1, v_1) \geq 0.$$

La cópula trivial es $\mathscr{C}(u_1, u_2) = u_1 u_2$: esta sirve como un caso límite de independencia completa para comparar cópulas "reales".

Un tipo especial de cópula es la *cópula Arquimediana*, la cual tiene la forma

$$\mathscr{C}(u_1, u_2) = \psi^{-1}\left(\psi(u_1) + \psi(u_2)\right). \tag{13.6.4}$$

La función $\psi(\cdot)$ se define sobre el intervalo $(0, 1)$, $\psi(1) = 0$, $\psi'(u) < 0$ y $\psi''(u) > 0$. Estas condiciones garantizan que la función ψ^{-1} sea doblemente diferenciable.

En Díaz y Morales [21, p. 225] se extiende la cópula Arquimediana para el caso multivariado ($\mathbf{u} \in \mathbf{I}^p$), así:

$$\mathscr{C}(\mathbf{u}) = \psi^{-1}\left(\psi(u_1) + \cdots + \psi(u_p)\right). \tag{13.6.5}$$

La cópula de *Clayton* se genera desde la cópula Arquimediana mediante $\psi(u) = (u^{-\nu} - 1)/\nu$, la forma de la cópula p-variante es

$$\mathscr{C}(\mathbf{u}) = \left\{\sum_{j=1}^{p} u_j^{-\nu} - (p-1)\right\}^{-1/\nu}.$$

La función de densidad es la correspondiente p-ésima derivada de $\mathscr{C}(\mathbf{u})$:

$$\partial^p \mathscr{C}(\mathbf{u})/\partial u_1 \cdots \partial u_p = \mathscr{C}(\mathbf{u})^{1+p\nu} \left\{\prod_{j=1}^{p-1}(1 + j\nu)\right\}\left(\prod_{j=1}^{p} u_l^{-\nu-1}\right).$$

En la función de sobrevida, la cópula puede ser dada en la fragilidad. Se empieza con

$$S(\mathbf{t}|z) = \exp\left\{-z\sum_{j=1}^{p} H_j(t_j)\right\}.$$

Ahora suponga que Z tiene distribución gama con parámetro de forma ν y parámetro de escala 1. La integral sobre la fragilidad z produce la función de sobrevida conjunta no condicional

$$S(\mathbf{t}) = (1+s)^{-\nu},$$

en la cual $s = \sum_{j=1}^{p} H_j(t_j)$; la distribución de Burr multivariada es un caso especial cuando H_j es de tipo Weibull. Las marginales son $S_j(t_j) = [1 + H_j(t_j)]^{-\nu}$, por lo tanto, $H_j(t_j) = S_j(t_j)^{-1/\nu} - 1$. Así, se puede escribir

$$S(\mathbf{t}) = \left\{\sum_{j=1}^{p} S_j(t_j)^{-1/\nu} - (p-1)\right\}^{-\nu},$$

la cual es una cópula tipo Clayton.

Un estudio más detallado de las cópulas se encuentra en Nelsen [56] y en Joe [35].

13.7. Modelos de fragilidad o vulnerabilidad

La comparación de historias de eventos y tiempos de sobrevida entre los miembros de una población puede advertir acerca de la heterogeneidad entre ellos, en relación con los factores de riesgo relevantes. La última fuente de variabilidad es conocida como vulnerabilidad, propensión, susceptibilidad o fragilidad (*frailty*). Así, en estudios clínicos, por ejemplo con la muerte o la recurrencia de una dolencia, algunos pacientes sobrevivirán o permanecerán saludables durante un tiempo largo a pesar de los factores de riesgo observables, mientras que algunos otros sobrevivirán tiempos más cortos de lo esperado. Esta heterogeneidad puede alterar la estimación del riesgo o de la tasa del evento de interés, lo que no representa el riesgo de cualquier individuo particular o subgrupo de la población [12, pp. 438-443]. La idea es que los individuos tienen diferente vulnerabilidad o fragilidad, y que aquellos cuya vulnerabilidad sea mayor, alcanzarán el evento de interés más pronto que los demás. Los modelos de vulnerabilidad se emplean también para hacer ajustes de sobredispersión en los datos de sobrevivencia.

Suponga que se consideran unas pocas subpoblaciones, las cuales se definen de acuerdo con el nivel de vulnerabilidad (exposición al riesgo). Para datos de sobrevivencia, la subpoblación más frágil tenderá a sufrir el evento más temprano, de tal forma que, con el paso del tiempo, la tasa de riesgo global descenderá para el grupo con la menor vulnerabilidad. Así, las tasas de riesgo para una subpoblación particular pueden declinar, aunque la tasa global se observe constante, y viceversa. Un ejemplo consiste en g subpoblaciones o subgrupos, cada uno con una tasa de riesgo $h_i(t)$, para $i = 1, 2, \ldots, g$. Las tasas de sobrevivencia para cada grupo vienen dadas por

$$S_i(t) = \exp[-\int_0^t h_i(u)du], \quad i = 1, 2, \ldots, g. \tag{13.7.1}$$

Si $p_i(0)$ denota la proporción inicial en la subpoblación $i = 1, 2, \ldots, g$, con $p_1(0) + p_2(0) + \cdots + p_g(0) = 1$, entonces la proporción que sobrevive en la cohorte que proviene de la subpoblación i en el tiempo t es

$$p_i(t) = \frac{p_i(0)S_i(t)}{\sum_{i=1}^{g} p_i(0)S_i(t)}, \quad i = 1, 2, \ldots, g. \tag{13.7.2}$$

La tasa de riesgo para la cohorte completa al tiempo t es

$$h_c(t) = \sum_{i=1}^{g} p_i(t)h_i(t). \tag{13.7.3}$$

Un modelo simple para tales subpoblaciones es el modelo de *permanencia-movilidad*. Por ejemplo, hay una subpoblación que tiene alto riesgo de divorcio, migración, cambio de empleo o enfermedad, y hay otra que tiene cero o bajo el respectivo riesgo. Si la tasa de riesgo para los grupos vulnerables es creciente, entonces el riesgo observado en la población completa puede primero ascender y luego decaer. De manera más general, se pueden modelar las variaciones no observadas en la fragilidad mediante el uso de mezclas discretas o mediante una densidad continua para la vulnerabilidad.

Un impacto de la no consideración de la heterogeneidad es que los efectos de las covariables pueden ser estimado con demasiada precisión. Así, por ejemplo, en análisis de mortalidad, si los riesgos no son homogéneos sobre los individuos, entonces el efecto de la edad sobre la mortalidad individual es mayor que sobre la mortalidad en población.

Un modelo de vulnerabilidad es un modelo de efectos aleatorios para datos de tiempo hasta la ocurrencia de un evento (de sobrevivencia), donde el efecto aleatorio tiene un efecto multiplicativo sobre la función de riesgo base. Se distinguen dos clases de modelos de vulnerabilidad: *los modelos con*

tiempo de sobrevivencia univariado y *los modelos de sobrevivencia multivariados* (por ejemplo, de riesgos competentes o de eventos recurrentes en el mismo individuo).

En el primer caso, una variable de tiempo de vida es empleada para describir la influencia de covariables no observables en el modelo de riesgos proporcionales. La variabilidad de los datos de sobrevivencia se divide en una parte que depende de los factores de riesgo, y que es teóricamente predecible, y una parte que es inicialmente no predecible, aunque toda la información relevante es conocida. La separación de estas dos fuentes de variabilidad trae como ventaja que la heterogeneidad puede explicar algunos resultados no esperados o dar una interpretación alternativa.

Para el segundo caso, cuando se consideran los tiempos de sobrevivencia multivariados en la forma de eventos recurrentes, uno de los objetivos de estos modelos es tener en cuenta la dependencia entre las observaciones sobre un mismo individuo. Una manera natural de modelar la dependencia entre los tiempos para eventos recurrentes es mediante la inclusión de una variable asociada con el efecto aleatorio; esta se conoce como vulnerabilidad.

13.7.1. Modelo de vulnerabilidad compartida para datos recurrentes

Las dos fuentes de variación consideradas arriba tienen ahora una interpretación diferente. La variación en vulnerabilidad no es una variación de grupo, sino una variación entre individuos, y la variación descrita por la función de riesgo no es una variación individual, sino una variación dentro de los individuos, la cual, para eventos recurrentes alternativamente, podría llamarse variación de Poisson. De esta forma, el supuesto de riesgos constantes es relevante para modelar la vulnerabilidad en datos de eventos recurrentes.

Para datos paralelos, el conjunto de individuos bajo riesgo decrece en cada tiempo que se presente el evento, mientras que para datos de eventos recurrentes, el conjunto bajo riesgo es constante durante el período de observación. Así, en datos paralelos, es importante observar los tiempos de los eventos, mientras que para datos de eventos recurrentes, la aproximación por vulnerabilidad permite variar el número de eventos, aunque el tiempo de observación es el mismo para todos los individuos. En una situación extrema de un período de observación muy largo, en el caso paralelo todos habrán muerto, por lo tanto, no hay variación en el numero de eventos

(muerte). En cambio, para el caso de eventos recurrentes, la variación será muy alta.

El valor de la vulnerabilidad Z es constante sobre el tiempo y común a los tiempos para los eventos observados sobre cada individuo; así, es el responsable de crear dependencia. Esta es la razón de la denominación de *compartidos*.

Notación

Los datos se describen como un proceso para cada individuo incluido en el estudio, los cuales se asume que son independientes. Cuando solo un tipo de evento es posible, los tiempos de los eventos para un individuo i se pueden ordenar $0 < T_{i1} < T_{i2} < \ldots$. Como se consideran modelos continuos, la probabilidad de que dos observaciones consecutivas coincidan es cero. Si el número de eventos es finito, dígase q, se define $T_{i,q+1} = T_{i,q+2} = \cdots = \infty$, en cuyo caso, los tiempos solo satisfacen que $0 \leq T_{i1} \leq T_{i2} \leq \ldots$. El número de eventos acumulados en el individuo i hasta el tiempo t se nota por $K(t)$; mientras que el proceso completo es denotado por $\{K(t)\}$, definido en el intervalo $[0, \infty]$, con $K(0) = 0$ y $K(\infty)$, que puede ser finito o infinito. Este es un proceso de conteo creciente sobre el tiempo, continuo por la derecha, con valores $\{0, 1, \ldots\}$.

El período de observación $(0, C]$ es finito y, por tanto, el número total de eventos observados es finito, $K = K(C)$. Es decir, el número de eventos K es aleatorio, $K \in \{0, 1, \ldots\}$.

13.7.1.1. Proceso de Poisson

Con el supuesto de que el riesgo de experimentar un evento es $\lambda(t)$, independientemente del estado actual, la distribución del número de eventos $(K(t) - K(v))$ en el intervalo $(v, t]$ es Poisson, con media $\int_v^t \lambda(u)du = \eta$. Así, las probabilidades son

$$P(K(t) - K(v) = j) = \eta^j \exp(-\eta)/j!. \tag{13.7.4}$$

Para el caso de riesgo constante, el proceso es llamado de Poisson homogéneo y el número total de eventos K es suficiente, donde $K = K(C)$ tiene una distribución de Poisson con media $C\lambda$. Esta es una familia exponencial, en la cual el parámetro canónico es $\log \eta$. Así, la varianza de esta distribución es igual a su media, η.

La verosimilitud de los tiempos observados, para el individuo i, viene dada por

$$\left\{\prod_{j=1}^{K}\lambda(T_j)\right\}\exp\left\{-\int_0^C \lambda(u)du\right\}. \tag{13.7.5}$$

En esta expresión se tiene la existencia de un término con el riesgo para cada evento observado en el individuo i y un término exponencial que describe el riesgo total integrado experimentado por el i-ésimo individuo.

Este modelo puede ser manejado con la función de riesgo general y con algunas covariables que actúen en forma proporcional al riesgo, como se contempla en el modelo de riesgos proporcionales dado por (10.3.1).

13.7.1.2. Modelos de vulnerabilidad constante con sobredispersión

Como se describió arriba, un modelo de vulnerabilidad compartida puede ser considerado como un modelo de efectos aleatorios con dos fuentes de variación. De una parte está la variación de grupo, descrita por la variable aleatoria Z (la vulnerabilidad), y de otra parte, se tiene la simple (o individual) variación aleatoria descrita por la función de riesgo, la cual se nota por $\mu(t)$.

En estos modelos se asume que todas los tiempos observados son independientes, dados los valores de las vulnerabilidades; es decir, es un modelo de independencia condicional. Se asume que el riesgo no cambia por los eventos ocurridos dentro de cada individuo. Así, condicionado al valor de la vulnerabilidad Z, se asume un proceso de Poisson, con riesgo $Z\mu(t)$. Condicionado a Z, la probabilidad es encontrada por inserción directa en la ecuación (13.7.5) y después integrando sobre Z. La contribución de verosimilitud basada en los eventos observados K, antes del tiempo t, es

$$P(K=k, T_1=t_1, \ldots, T_k=t_k) = (-1)^k\left\{\prod_{j=1}^{k}\mu(t_j)\right\}\mathcal{L}^{(k)}\left(\int_0^t \mu(u)du\right), \tag{13.7.6}$$

donde $\mathcal{L}(s) = E\exp(-sZ)$ es la transformada de Laplace de la variable aleatoria Z y $\mathcal{L}^{(k)}$ su derivada de orden k. Para obtener la distribución del número total de eventos (K) durante $(0, t]$, se integra $t_1, \ldots, t_K$ sobre $(0, t)$, sujeto a la restricción $0 \leq t_1 \ldots \leq t_K \leq t$. Esta integral puede evaluarse integrando sobre el conjunto producto $[0, t]^K$ y dividiendo por $K!$. La distribución de K es de Poisson con media $Z\int_0^t \mu(u)du$, condicionalmente a

Z; e incondicionalmente, la distribución está dada por

$$P(K=k) = (-1)^k \left\{ \int_0^t \mu(u)du \right\}^k \mathscr{L}^{(k)}\left(\int_0^t \mu(u)du \right)/k!. \qquad (13.7.7)$$

La distribución de los tiempos para los eventos condicionados sobre el número total de eventos, K, es una distribución con densidad

$$f(t_1 \ldots, t_k|K) = \mathbf{1}\{0 \le t_1 \ldots \le t_K \le t\} \prod_{j=1}^{K} \{\mu(t_j)/\int_0^t \mu(u)du\} K!. \qquad (13.7.8)$$

Se sigue que la distribución de la vulnerabilidad solo entra en la distribución del número total de eventos (ecuación 13.7.7), y esta depende del riesgo a través del riesgo integrado sobre el intervalo completo de observación. La distribución de los tiempos, condicionada al número de eventos, depende solo del riesgo relativo (esto es, relativo al riesgo integrado), dado en la ecuación (13.7.8). Esto significa que el número de eventos en cualquier intervalo dado, $(v, t]$, es una distribución binomial con parámetro de probabilidad $\int_v^t \mu(u)du / \int_0^t \mu(u)du$, condicionado sobre el número total de eventos.

En particular, cuando $\mu(t)$ es constante, se obtiene la expresión

$$P(K=k) = (-1)^k (t\mu)^k \mathscr{L}^{(k)}(t\mu)/k!, \qquad (13.7.9)$$

donde la distribución de los tiempos, condicionada al número total de eventos, es la distribución de las observaciones ordenadas para K variables aleatorias independientes con distribución uniforme $(0, t]$.

La tasa de riesgo de un evento al tiempo t, condicionada a la historia hasta ese momento, puede se obtenida a partir de la ecuación (13.7.6). Esto quiere decir que se puede actualizar la distribución de la vulnerabilidad. Supóngase que han ocurrido k eventos hasta el tiempo t, por lo que el riesgo de un evento adicional es

$$(-1)\mu(t)\mathscr{L}^{(k+1)}\left(\int_0^t \mu(u)du \right)/\mathscr{L}^{(k)}\left(\int_0^t \mu(u)du \right). \qquad (13.7.10)$$

Se observa que el riesgo es independiente de los tiempos asociados con los eventos previos; únicamente el número de eventos entra en la expresión anterior. En consecuencia, este modelo es un modelo de Markov. Otro significado de la ecuación (13.7.10) es que toma la forma $\lambda_k(t)$.

13.7.1.3. Modelos de regresión

Puede ser de interés la inclusión de covariables, X, con el fin de evaluar el efecto de estas variables y explorar si ellas explican parte de la variación entre procesos. Cuando las variables son conocidas se incluyen, generalmente, en un modelo de riesgo proporcional, con el propósito de verificar el efecto de estas sobre el riesgo. De esta manera, el modelo que describe el riesgo de experimentar un evento al tiempo t para el i-ésimo individuo es

$$Z_i \exp(\beta^T x_i)\mu_0(t), \tag{13.7.11}$$

el cual es condicional sobre los Z_i. x_i es una covariable constante o común en el tiempo de observación. Los efectos de estas covariables evidenciarán la variación entre individuos. En un experimento *cross-over*, la covariable podría describir los tratamientos y, de esta forma, serían dependientes del tiempo, $x_i(t)$.

Un modelo alterno al modelo dado en (13.7.11) consiste en emplear el tiempo entre eventos consecutivos (*gap times*, en inglés). Cuando el individuo ha experimentado j eventos, el riesgo de un evento es $Z_i \exp(\beta^T x_i)\varphi_0(t-T_j)$. Este es un proceso de renovación donde la distribución de los tiempos de interocurrencia de eventos difiere entre individuos como una consecuencia de la vulnerabilidad individual.

Hay casos en los cuales los eventos recurrentes no son independientes de un evento terminal. Por ejemplo, la recurrencia de ataques cardíacos frecuentemente incrementa el riesgo de muerte, lo cual imposibilita que se presente cualquier otro subsecuente evento recurrente. Liu *et al.* [50] proponen un modelo de vulnerabilidad compartido y conjunto para eventos recurrentes y un evento terminal. Otra aplicación de los modelos de vulnerabilidad es el modelamiento conjunto de datos longitudinales y tiempo para un evento. Ratcliffe *et al.* [60] presentan un método para el modelamiento conjunto de datos longitudinales y de sobrevivencia mediante los modelos de vulnerabilidad.

13.7.1.4. Estimación

La estimación se hace dependiendo de que los datos correspondan a tiempos exactos, períodos de conteo o a intervalos de conteo. En cada uno de estos casos, la función de verosimilitud es diferente. Para tiempos exactos, se tienen observaciones en tiempo continuo; en este caso, una función de riesgo no paramétrica y la presencia de censura es una posibilidad. Para períodos de conteo, la distribución es un modelo paramétrico para las obser-

vaciones sobre enteros no negativos. La situación para intervalos de conteo es similar a la de períodos de conteo, pero el conteo es multidimensional.

13.7.1.5. Una aplicación

El número de reclamos de pólizas de seguros automovilísticos por policía, sobre un período fijo de tiempo, corresponde a datos del período de conteo. La siguiente tabla contiene los datos originales y ajustados mediante varios modelos de Poisson con sobredispersión.

Tabla 13.9. Número de reclamos en pólizas

Reclamos	**Policías**	**Poisson**	**Bin. Neg.**	**Inv. Gaus.**
0	103704	102629.6	103723.6	103710.0
1	14075	15922.0	13990.0	14054.7
2	1766	1235.1	1857.1	1784.9
3	255	63.9	245.2	254.5
4	45	2.48	32.29	40.42
5	6	0.08	4.24	6.94
6	2	0.002	0.557	1.258
≥ 7	0	0.00004	0.084	0.295

El ajuste se hace para distribuciones tales como Poisson, binomial negativa, inversa gausiana, entre otras. El cálculo de los respectivos coeficientes de variación advierte acerca de una posible sobredispersión.

Los parámetros de una distribución de Poisson se estiman mediante la media empírica y corresponde a 0.1551 ($EE = 0.0011$), y la varianza empírica total es 0.1793. El coeficiente de variación empírico para Z es 1.003, lo cual sugiere la presencia de una sobredispersión. El logaritmo de la verosimilitud es igual a -55108.46.

Para la binomial negativa, los parámetros estimados son: $\hat{\delta} = 1.033(ee = 0.044)$ y $\hat{\theta} = 6.565(ee = 0.286)$, y el logaritmo de la verosimilitud es igual a -54615.32, que corresponde a un incremento cerca de 500, y así muestra un mejor ajuste. El coeficiente de variación de Z estimado es de 0.984.

Los parámetros de la inversa gausiana son estimados como: $\hat{\delta} = 0.2784(ee = 0.0063)$ y $\hat{\theta} = 3.220(ee = 0.148)$, y el logaritmo de la verosimilitud es igual a -54609.76, lo cual muestra un ajuste mejor que el mostrado por la binomial negativa.

Los datos ajustados por cada una de estas distribuciones se muestran en la tabla anterior.

Algunas otras aplicaciones sobre cáncer de mama y ataques de epilepsia se encuentran en Hougaard [33, pp. 336-338].

13.7.2. Análisis de eventos recurrentes

En muchos estudios científicos, el interés frecuentemente se centra en el proceso que generan eventos sobre una misma unidad de manera repetida en un periodo de tiempo. Así, hay unidades que pueden experimentar el mismo evento varias veces durante la duración del estudio $[0, \tau)$. Tales procesos proveen datos que se denominan *datos de eventos recurrentes*.

En eventos recurrentes hay principalmente dos tipos de tiempos. De una parte el tiempo, desde un origen definido (tiempo calendario) hasta que el evento ocurra; de otra, el tiempo entre la ocurrencia de dos eventos consecutivos sobre la misma unidad. En la construcción de modelos para este tipo de datos, se debe considerar e incluir en el modelo alguno de estos tipos de tiempo.

Una revisión de los métodos y modelos estadísticos para datos de eventos recurrentes se presentan en Cai y Schaubel [8] y en Cook y Lawless [13]; allí se describen los principales modelos, junto con los supuestos y propiedades relevantes. Ellos consideran métodos paramétricos, no paramétricos y semiparamétricos, tanto para un solo tipo de evento como para varias clases de eventos.

Para los propósitos de este trabajo, se hace una revisión de los principales modelos para eventos recurrentes, los cuales se generalizarán en el esquema de trabajo de esta tesis.

Se revisa el modelamiento para datos de eventos recurrentes desde el trabajo pionero en esta área de Gail [27].

13.7.2.1. Gail, Santner y Brown (1980)

Estos autores proponen métodos para comparar dos tratamientos sobre ratones que han tenido recurrencia de tumores del mismo tipo. Ellos presentan métodos basados en el tiempo entre apariciones consecutivas (interocurrencia) del tumor. En los ensayos sobre cáncer, algunas veces resulta la reaparición de tumores, por ejemplo, K tumores en un animal. Así, los tiempos observados corresponden a valores de las variables aleatorias $T_1 < T_2 < \cdots < T_K$.

Se podría comparar el número medio de tumores por animal en dos grupos de animales experimentales si cada uno de los animales tiene el mismo período de seguimiento. Los autores no asumen la presencia de

censura a derecha. Se puede comparar el tiempo hasta el k-ésimo tumor $k = 1, \ldots, K$. A partir de las respectivas tablas de sobrevida, se puede estimar la proporción de animales que tienen menos de k tumores en un tiempo τ y usar esta proporción como una base para comparar los dos tratamientos.

Los autores exploran algunos métodos semiparamétricos para combinar información a partir de tiempos interocurrencia consecutivos, $Z_i = T_i - T_{i-1}$, y obtienen estadísticas resumen para el efecto de tratamientos. Además, exploran una metodología en la cual es posible combinar la evidencia basada en los tiempos T_i en lugar de los tiempos interocurrencia Z_i. Bajo la hipótesis de no efecto de tratamientos, la distribución conjunta de las estadísticas de orden $T_1 \leq T_2 \leq \cdots \leq T_K$ es la misma para todos los animales. El tiempo de censura C para cualquier animal se asume independiente de la sucesión $T_1,\ T_2, \ldots$, donde K es simplemente el mayor entero tal que $T_K < C$. Bajo esta formulación, la distribución condicional del i-ésimo tiempo interocurrencia Z_i, ya que $T_1 = t_1, \ldots, T_{i-1} = t_{i-1}$, está dada por

$$pr(Z_i \geq z | t_1, t_2, \ldots, t_{i-1}) = exp\left\{ - \int_0^z h(\tau | i, t_1, t_2, \ldots, t_{i-1}) d\tau \right\}, \tag{13.7.12}$$

donde h es una función de riesgo no negativa. Los autores simplifican (13.7.12) bajo el supuesto de que $T_1,\ T_2, \ldots$ es una sucesión de Markov, así:

$$pr(Z_i \geq z | t_1, t_2, \ldots, t_{i-1}) = exp\left\{ - \int_0^z h(\tau | i, t_{i-1}) d\tau \right\}. \tag{13.7.13}$$

Basados en la premisa de que el cáncer proviene de la proliferación de una serie de células "copias" de una misma, para el sitio m, en el modelo, se asume que cada animal tiene m sitios afectados con un tumor, es decir, células cuyo tiempo para alcanzar el estado de tumor, X_i, $i = 1, 2, \ldots, m$, es independiente e idénticamente distribuido. El tiempo común para la detección de tumores es denotado por:

$$S(y) = pr(X_i \geq t) = exp\left\{ - \int_0^t \lambda(u) du \right\}, \tag{13.7.14}$$

donde $\lambda(\cdot)$ es el riesgo común.

Por los resultados conocidos sobre estadísticas de orden, se tiene que

$$pr(Z_i \geq z | t_1, t_2, \ldots, t_{i-1}) = \left\{ S(t_{i-1+z} / S(t_{i-1}) \right\}^{m-i+1}. \tag{13.7.15}$$

Se sigue que la función de riesgo para el tiempo interocurrencia i es

$$h(z | i, t_1, t_2, \ldots, t_{i-1}) = (m - i + 1) \lambda(t_{i-1+z}), \tag{13.7.16}$$

la cual satisface el supuesto de Markov. En general, cualquier modelo, cuya distribución de i-ésimo tiempo interocurrencia Z_i dependa de $Z_1, Z_2, \ldots, Z_{i-1}$ únicamente a través de la suma de los $Ti-1$, satisface también (13.7.15).

Si el modelo para la función de riesgo en el m-ésimo sitio fuera $\lambda(t)$ en el grupo 2, y $exp(\alpha)\lambda(t)$ en el grupo 1, entonces (13.7.16) implica que el riesgo condicional en el tiempo interocurrencia para el grupo 1 y el grupo 2 es $exp(\alpha)h(z|i, t_{i-1})$ y $h(z|i, t_{i-1})$, respectivamente. Por tanto, la siguiente adaptación del modelo de riesgos proporcionales, de acuerdo a Crowder [17], es

$$exp(\alpha_i)h(z|i, t_{i-1}) \qquad o \qquad h(z|i, t_{i-1}), \tag{13.7.17}$$

dependiendo si el animal pertenece al grupo 1 o 2, respectivamente.

13.7.2.2. Andersen y Gill (1982)

El modelo de Andersen y Gill (en adelante, AG) es muy cercano a un modelo de regresión de Poisson. En el modelo de AG, la intensidad del proceso para la i-ésima unidad, dadas las covariables $W(t)$, es

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{W_i'(t)\beta\}. \tag{13.7.18}$$

Este es un modelo para eventos recurrentes, el cual es una extensión natural del modelo de regresión de Cox. La diferencia entre el modelo de Cox y el modelo AG está en la definición del indicador de riesgo $Y_i(t)$. En el modelo de Cox, la i-ésima unidad termina de estar en riesgo una vez que al evento le haya ocurrido por primera vez y, por lo tanto, $Y_i(t)$ pasa de tomar el valor 1 a tomar el valor 0, mientras que en el modelo de AG, para eventos recurrentes $Y_i(t)$, permanece igual a 1 cuando el evento ocurre sobre la misma unidad i.

El modelo de AG es apropiado para situaciones en las que los eventos observados sobre una misma unidad se pueden asumir mutuamente independientes. El modelamiento es desarrollado conforme a un proceso de Poisson dinámico.

13.7.2.3. Prentice, Williams y Peterson (1981)

Ellos consideran dos clases generales de modelos de regresión para eventos recurrentes (en adelante, PWP), los cuales relacionan el riesgo como una función de intensidad con covariables y la historia de falla. Los modelos consideran e incluyen el tiempo desde el origen del estudio hasta la ocurrencia de cada evento y el tiempo interocurrencia, respectivamente. Ambos modelos son estratificados de riesgos proporcionales, lo cual significa

que la función de intensidad puede variar de un evento a otro, mientras que en el modelo de AG se asume que todos los eventos son idénticos.

Los métodos se pueden considerar como una generalización de las técnicas de análisis de datos de sobrevida, en las cuales la función de riesgo es modelada, sobre una unidad, más allá de la ocurrencia del primero, el segundo y los siguientes eventos. De otra forma, una unidad no está en riesgo para el k-ésimo evento hasta que haya experimentado el $k-1$-ésimo evento.

Sean $W(u) = W_1(u), \ldots, W_p(u)$ un vector de covariables, para una unidad bajo estudio, la cual está bajo observación en el tiempo $u \geq 0$, y donde $W(t) = \{w(u) : u \leq t\}$ es el correspondiente proceso de covariables hasta el tiempo t. Similarmente, sea $N(t) = \{n(u) : u \leq t\}$, donde $n(u)$ es el numero de eventos o episodios antes del tiempo u. La función de riesgo o de intensidad al momento t es definida como una tasa instantánea de falla al momento t, y dadas las covariables y el proceso de conteo en el tiempo t, es

$$\lambda\{t|N(t), W(t)\} = \lim_{\Delta t \to 0} pr\{t \leq T_{n(t)+1} < t + \Delta t|N(t), W(t)\}/\Delta t. \tag{13.7.19}$$

Siguiendo a Crowder [17] y a Therneau y Grambsch [66], se puede escribir (13.7.19) como el producto entre una función arbitraria y una función exponencial en las covariables. Los autores presentan dos tipos de funciones de línea base, una en función del tiempo desde el inicio del estudio hasta el momento, t, y otra desde el evento inmediatamente anterior, $t - t_{n(t)}$. Además, parece conveniente permitir que la forma de la función de riesgo dependa del número de eventos anteriores y posiblemente de otras características; esto se condensa en $\{N(t), W(t)\}$. Así, se dispone de dos modelos de riesgo parcialmente paramétricos, estos son, respectivamente,

$$\lambda\{t|N(t), X(t)\} = \lambda_{0s}(t) \exp\{W'(t)\beta_s\} \tag{13.7.20}$$

y

$$\lambda\{t|N(t), X(t)\} = \lambda_{0s}(t - t_{n(t)}) \exp\{W'(t)\beta_s\}, \tag{13.7.21}$$

donde, en cada caso, $\lambda_{0s}(,) \geq 0$ $(s = 1, 2, \ldots)$ son funciones de intensidad base arbitrarias. La variable de estratificación $s = s\{N(t), W(t), t\}$ puede variar como una función del tiempo para una unidad dada, y β_s es un vector columna de los coeficientes de regresión estratificados. Un caso especial de variable de estratificación es $s = n(t) + 1$, para el cual una unidad se muda al estrato k después de que le ha ocurrido el $k-1$-ésimo evento, y permanece allí hasta el siguiente evento $(k + 1)$ o queda bajo censura.

La función exponencial en (13.7.20) o (13.7.21) puede cambiarse a otra función no negativa. Note también que el modelo de Cox puede ser considerado como un caso especial de (13.7.20), para el cual $\lambda_{0s}t = \lambda_0(t)$ para todo s. Gail *et al.* [27] consideran los dos casos especiales de (13.7.21) con estratos definidos como $s = n(t) + 1$.

13.7.2.4. Wei, Lin y Weissfeld (1989)

Wei, Lin y Weissfeld (en adelante, WLW) proponen métodos semiparamétricos para analizar tiempos de falla multivariados. Modelan la distribución marginal de cada uno de los tiempos de falla mediante un modelo de Cox. Sea T_{ij} el tiempo para la j-ésima ocurrencia del evento sobre la i-ésima unidad, con $j = 1, \ldots, \mathcal{K}$ y $i = 1, \ldots, n$; donde $\mathcal{K}$ es el máximo número de eventos observado en los datos.

Sea $W_{ij} = (W_{ij,1}(t), \ldots, W_{ij,p}(t))'$ un vector $p\times$ de covariables para la i-ésima unidad en el tiempo $t \geq 0$ respecto al j-ésimo evento.

En los modelos de AG y PWP no se considera ni se incluye la estructura de dependencia intraunidad debida a las observaciones repetidas del evento. Los modelos de WLW son propuestos para modelar funciones de riesgo marginal mediante funciones de intensidad condicionadas a un proceso de conteo $N_i(t)$.

Para la j-ésima ocurrencia del evento sobre la i-ésima unidad, la función de riesgo $\lambda_{ij}(t)$ se asume que toma la forma

$$\lambda_{ij}(t) = Y_{ij}\lambda_{0j}(t)\exp\{W_{ij}'(t)\beta_j\}. \tag{13.7.22}$$

Note que este modelo estratifica por el orden de ocurrencia del evento. Como $\mathcal{K}$ es el número máximo de eventos sobre alguna de las unidades, es natural que si a una unidad le ocurre un número de eventos H menor que $\mathcal{K}$, esta tendrá valores faltantes sobre la ocurrencia del evento después de la H-esima ocurrencia.

Este modelo permite desarrollar un análisis por separado para cada estrato j y para cada interacción de estrato y covariable; esto se lee desde los subíndices j escritos en el modelo (13.7.22). El indicador de riesgo, Y_{ij}, es igual a 1 hasta la ocurrencia del j-ésimo evento, a menos que algunas condiciones externas causen censura sobre la unidad. Ninguna estructura particular de dependencia entre los tiempos de falla en cada unidad es impuesta. Los parámetros de regresión son estimados mediante la respectiva función de verosimilitud parcial.

13.7.2.5. Chang y Wang (1999)

Chang y Wang proponen un modelo de riesgo semiparamétrico para tiempos de eventos recurrentes (en adelante, CW) a través de un modelo de regresión condicional utilizando los modelos de Cox y PWP. El modelo es como sigue:

$$\lambda_{ij}(t|\mathcal{H}(t), Z(t)) = \lambda_{j0}(t - t_{j-1}) \exp\{\beta' W_{i1}(t) + \gamma_j W_{i2}(t)\}, \quad (13.7.23)$$

donde $\mathcal{H}(t) = \{T_1, \ldots, T_{j-1} : T_1 < T_2 < \cdots < T_{j-1} < t \leq T_j\}$ es la historia del evento hasta el tiempo t.

El modelo contiene el parámetro estructural, β, y el parámetro, γ_j, asociado con el episodio específico, los cuales corresponden a los efectos asociados con las covariables $W_{i1}(t)$ y $W_{i2}(t)$, respectivamente. Si $\beta = 0$, el modelo (13.7.23) se reduce al modelo de PWP expresado en (13.7.21). Este modelo es útil cuando el interés se focaliza en episodios u ocurrencias específicas del evento.

La estimación de los parámetros es obtenida a partir de la función de puntaje para β sujeta a que γ_j es fijo, y recíprocramente, a partir de la función de puntaje para γ_j a condición de que se tiene β. En ambos casos, las funciones de puntaje se obtienen de la función de verosimilitud parcial que los parámetros y los datos definen.

13.8. Modelos multiestado

Una unidad puede alcanzar diferentes estados en tiempos distintos. El interés se centra en modelar el tiempo hasta que la unidad alcanza cada uno de varios estados. En los procesos de degradación se presentan movimientos progresivos a través de una sucesión de estados. Otro ejemplo es el estado de salud de un paciente, el cual varía con el tiempo, con la esperanza de mejorar, aunque se presentan contratiempos periódicos. Estos tiempos a menudo pueden ser censurados por intervalo, por ejemplo, cuando ocurren cambios de estado entre visitas consecutivas al odontólogo.

Los modelos multiestado son los modelos más utilizados para describir el desarrollo de datos de tiempos a eventos de tipo longitudinal. Un modelo multiestado se define como un modelo para un proceso estocástico, donde una unidad en cualquier punto del tiempo ocupa uno de los estados de un conjunto de estados discretos i finito. En medicina, los estados pueden describir las condiciones de salud o desarrollo de una enfermedad de paciente, como sano, enfermo, enfermo con complicaciones y muerto. Un cambio

de estado se denomina transición. Por ejemplo, la aparición de una enfermedad, la ocurrencia de complicaciones y la muerte constituyen un proceso con estos tres estados.

La estructura del modelo de estado especifica qué estados y qué transiciones de un estado a otro son posibles y observables. Un modelo estadístico completo especifica la estructura del estado y la forma de la función de riesgo para cada transición posible. En este capítulo se hace una descripción o arquitectura de los modelos multiestado empleados para modelar datos de sobrevida multivariados y múltiples; la palabra *arquitectura* se emplea porque se deben mostrar los diferentes estados alcanzables por la unidad y las posibles conexiones o transiciones entres estos. La estructura del estado no es única. Al elegir la "mejor estructura", se pueden hacer los supuestos necesarios para que el modelo sea claro y simplifique los cálculos (parsimonia), pero que siempre se ajuste a la realidad contenida en los datos y no al revés. Se destacan algunos casos especiales como estructuras de estado estándar. En la figura 13.1 se muestran ejemplos de tales tipologías o estructuras (arquitecturas) de modelos.

En esta sección se hace una descripción de cómo los modelos multiestado se pueden emplear para modelar datos de sobrevida multivariados y múltiples. El tratamiento de estos modelos requiere de particular atención a la consideración e inclusión de la posible dependencia entre los eventos considerados. Además, los modelos multiestado consideran todos los datos como longitudinales (indexados en el tiempo), esto los hace menos útiles para mediciones repetidas y requieren el patrón de censura para que los datos en paralelo sean homogéneos.

El modelo de mortalidad (1), con estados "vivo" y "muerto", es el modelo multiestado más simple posible. De hecho, es tan simple que no hay ventaja en ponerlo en el marco de múltiples estados.

El modelo de riesgos competentes (2), con estados "vivo" y "muerto" debido a la causa j, con $j = 1, \ldots, k$, es un modelo apropiado para datos con causa del evento (en este caso, la muerte).

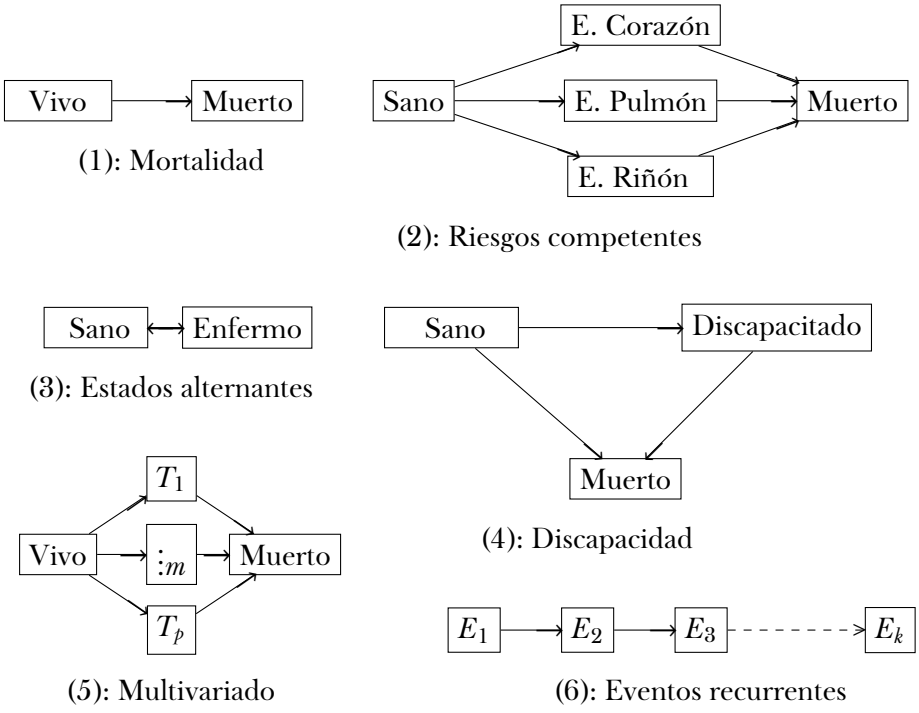

Figura 13.1. Tipologías de modelos multiestado.

Apéndice

A

Procedimientos básicos con R

$$S(t) = P(T > t) = \int_t^\infty f(u)du = 1 - F(t)$$

R es un sistema para realizar cálculos y gráficos estadísticos, cuenta con un poderoso lenguaje de programación orientado a objetos, un entorno para la ejecución de comandos, un depurador de código de programación, el acceso a determinadas funciones del sistema y la capacidad de ejecutar programas almacenados en archivos (*scripts*). R fue escrito por Ross Ihaka y Robert Gentleman en el Departamento de Estadística de la Universidad de Auckland, en Auckland, Nueva Zelanda. Además, un gran grupo de personas ha contribuido enviando código de programación e informes de error. R es un *software* libre, se puede descargar desde www.r-project.org/, donde se consigue en versiones para los sistemas operativos Windows®, Linux® y Macintosh®.

En este apéndice se explica cómo hacer algunas tareas usuales que se requieren en el análisis de datos, cómo efectuar cálculos de probabilidades y cuantiles y cómo leer datos externos, ordenarlos y transformarlos.

A.1. Lectura de datos externos

Cuando se analizan datos es usual que la digitalización de estos se haga a una hoja de cálculo tipo Excel o Calc de Libre Office; estos paquetes ofrecen más comodidad para este tipo de tareas. En esta sección se estudia cómo leer en R datos que se encuentran en archivos externos.

A.1.1. El directorio de trabajo

Una fuente de error frecuente para el que está iniciándose en R es olvidar fijar el *directorio de trabajo*, que es el directorio donde R buscará, por defecto, los archivos a leer. Si el archivo de datos que se pretende almacenar en la memoria no se encuentra en el directorio de trabajo, hay que especificar la ruta exacta del archivo, para que la lectura sea exitosa. Para no tener que digitar la ruta completa de ubicación de los archivos, se recomienda crear, para cada proyecto, un directorio donde se colocarán todos los archivos de datos y código asociados. Esto guardará los datos del proyecto como funciones creadas por el usuario o archivos de código.

La función que permite conocer cuál es el directorio de trabajo actual es `getwd()`.[1] Al ejecutar esta función, R devuelve la ruta del directorio de trabajo, como se muestra a continuación (la salida cambiará de acuerdo con la configuración del sistema).

[1] *get working directory.*

```
> getwd()
[1] "/home/mario"
```

Para cambiar el directorio de trabajo se usa la función `setwd()`.[2] Por ejemplo, si se desea fijar la carpeta `/home/mario/proyecto` como directorio de trabajo, se procede como sigue:

```
> setwd("/home/mario/proyecto")
```

A.1.2. Lectura de datos desde un archivo de texto

Cuando se tienen los datos en un archivo de texto plano, se cuenta con la función `read.table()`. Este tipo de archivo se reconoce por la extensión, que usualmente es .txt o .dat. Suponga que se desea leer este archivo[3]

```
height weight
  58    115
  59    117
   .      .
   .      .
   .      .
  71    159
  72    164
```

el cual se tiene guardado en el directorio de trabajo con el nombre `archivo.txt`. El comando es

```
datos<-read.table("archivo.txt",header=TRUE)
```

De esta forma, se crea el objeto `datos`, que contiene la información almacenada en `archivo.txt`. La opción `header=TRUE` es indispensable porque el archivo tiene, en su primera fila, los nombres de las columnas.

A.1.3. Lectura de datos desde un archivo CSV

Un archivo CSV es un archivo de texto en el cual los datos están separados por comas. La ventaja que tiene trabajar con este formato es que todas las hojas de cálculo son capaces de crearlos y leerlos. Es frecuente que el dueño

[2] En la versión para Windows, además del comando, el usuario cuenta con el menú Archivo > Fijar directorio de trabajo.

[3] *Data frame* `women` del paquete base.

de la información tenga sus datos organizados en una hoja de cálculo de Open Office (.odt) o Excel (.xls). Para importar dichos datos a la memoria de R, podemos crear un archivo con formato CSV y leerlo con la función `read.csv()` o con `read.csv2()`. Suponga que deseamos leer los siguientes datos, guardados en el directorio de trabajo con el nombre `ejemplocsv.csv`:

```
tiempo, momento, resp
 1, 0 ,95
 2, 0 ,93
 .  .  .
 .  .  .
 .  .  .
19, 1 ,7
20, 1 ,3
```

Note que las columnas están separadas por comas. La lectura de estos datos se hace mediante el comando

```
read.csv("ejemplocsv.csv")
```

Si en lugar de la coma tenemos un punto y coma separando las columnas, la función apropiada es `read.csv2()`. Por ejemplo, si tenemos los mismos datos anteriores, pero separados por punto y coma en un archivo ubicado en el directorio de trabajo, con nombre `ejemplocsv2.csv`,

```
tiempo ; momento ; resp
 1; 0 ;95
 2; 0 ;93
 .  .  .
 .  .  .
 .  .  .
19; 1 ;7
20; 1 ;3
```

la lectura se hace mediante el comando

```
read.csv2("ejemplocsv2.csv")
```

La diferencia entre `read.csv()` y `read.csv2()` radica esencialmente en el caracter usado como separador de campos, como se ilustra en el ejemplo anterior, y en el usado para el separador decimal. La primera usa `dec=","` mientras que la segunda usa por defecto `dec="."`. En realidad, el segundo archivo de datos se puede leer con el comando

```
read.csv("ejemplocsv2.csv",sep = ";")
```

A.1.4. Lectura de datos desde un archivo de Excel

Existen varias formas de leer un archivo en formato de Excel (.xls). Aquí se ilustra cómo hacerlo con las librerías `RODBC` y `gdata`. Suponga que tenemos el archivo `ejemplo.xls` con varias hojas y que los datos de interés se encuentran en la hoja 3. La lectura, mediante `RODBC`, se hace con el siguiente código:

```
library(RODBC) #debe estar instalada previamente
archivo<-odbcConnectExcel("ejemplo.xls")
tabla<-sqlFetch(archivo,"Hoja3")
close(archivo)
```

De esa forma se crea el objeto `tabla`, que contiene los datos de la hoja 3. La orden `close(archivo)` es para cerrar la conexión con el archivo. Mientras la conexión esté activa no es posible hacer cambios en el archivo `ejemplo.xls`. La lectura de la misma hoja del archivo, pero usando la librería `gdata`, se indica a continuación:

```
library(gdata) #debe estar instalada previamente
tabla<-read.xls("ejemplo.xls",sheet=3)
```

La función `read.xls()` trabaja convirtiendo, de manera temporal, el archivo de excel a un archivo CSV.

A.2. Selección y transformación de datos

Una vez cargado un conjunto de datos en la memoria de R, los pasos siguientes son la selección de un subconjunto de ellos, la creación de nuevas variables a partir de las antiguas y la obtención de algunas estadísticas básicas, bien sea en forma global o de acuerdo con algún criterio. En esta sección se ilustran, a través de ejemplos, los comandos básicos para realizar estas tareas. Para los ejemplos se usará la base de datos `survey` de la librería `MASS`, que se carga a la memoria con los siguientes comandos

```
library(MASS)
data(survey)
```

Para imprimir las primeras y las últimas 6 filas de los datos, se usan los siguientes comandos:

```
head(survey)
tail(survey)
```

Las estadísticas descriptivas básicas se obtienen con

```
summary(survey)
```

Cuando la variable es de caracteres o es un factor, el resumen consta de una lista de los niveles del factor acompañada de la frecuencia de cada nivel. Se observa que algunas variables tienen datos faltantes. Para eliminar del archivo las filas con datos faltantes, se procede como sigue:

```
datos<-na.omit(survey)
summary(datos)
```

De esta forma, creamos el objeto `datos` que no contiene datos faltantes (`NA`).

A.2.1. Creación de nuevas variables

Suponga que se requiere crear una nueva columna, a partir de las existentes en `datos`, mediante una transformación. Por ejemplo, para crear una columna que contenga el logaritmo natural de `Height`, procedemos así:

```
datos$lHeight<-log(datos$Height)
```

Otras funciones de uso frecuente son `sqrt()` (raíz cuadrada), `exp()` (exponencial), `abs()` (valor absoluto) y `sin()` (seno).

A.2.2. Selección de subconjuntos de un marco de datos

A continuación, se explica cómo usar el sistema de indexación de R para la selección de filas y columnas de un marco de datos.

A.2.2.1. Selección de columnas

El comando

```
datos[,1:4]
```

permite seleccionar las primeras 4 columnas del archivo `datos`. Note el uso de la coma, la cual separa filas de columnas (`[filas,columnas]`). En este caso, no hay ningún criterio para las filas; por lo tanto, se seleccionan todas. Si se quiere seleccionar las columnas numéricas, que están en las posiciones 2, 3, 6, 10 y 12, procedemos así:

```
datos[,c(2,3,6,10,12)]
```

Las columnas también se pueden seleccionar por sus nombres, como se indica en el siguiente código:

```
datos[,c("Wr.Hnd","NW.Hnd","Pulse","Height","Age")]
```

Para seleccionar las columnas que no son numéricas, dado que se conoce cuáles son las numéricas, se usa el siguiente código (note el uso del signo menos):

```
datos[,-c(2,3,6,10,12)]
```

La siguiente forma es interesante, porque la selección de las columnas es automática. Si se quiere seleccionar todas las variables numéricas (sin hacer una lista exhaustiva de ellas o de sus posiciones), se procede así:

```
numericas<-sapply(datos, is.numeric)
datos[,numericas]
```

La función `sapply` entrega `TRUE` donde hay una columna numérica y `FALSE` en caso contrario. El objeto `numericas` es lo que se conoce como un vector de tipo lógico. Mediante el comando `datos[,numericas]`, R selecciona las columnas cuya posición corresponda a un `TRUE` en el vector lógico `numericas`. Para seleccionar las columnas que no sean numéricas, se usa la orden

```
datos[,!numericas]
```

El signo de exclamación se usa en R para la negación lógica.

A.2.2.2. Selección de filas

El comando

```
datos[1:7,]
```

selecciona las 7 primeras filas del archivo `datos`. Si se desea seleccionar solo las 7 primeras filas de las columnas numéricas, se procede así:

```
datos[1:7,numericas]
```

El comando

```
datos[c(3,5,1,4,9), ]
```

selecciona las filas 3, 5, 1, 4 y 9, en ese orden. Una situación que se presenta a menudo es la selección de individuos (filas) de acuerdo con un criterio; por ejemplo, seleccionar a los hombres o seleccionar a las mujeres que fuman. A continuación, se muestran ejemplos concretos:

```
datos[datos$Sex=="Male", ]
```

selecciona todos los individuos hombres. El siguiente comando selecciona a las mujeres que nunca fuman (note el uso del operador lógico &):

```
datos[Sex=="Female" & Smoke=="Never",]
```

A.2.2.3. Ordenar una base de datos

A veces es necesario ordenar una base de datos por algún criterio. Aquí se explica cómo hacerlo.

```
datos[order(datos$Age),]
```

ordena el archivo datos por la columna edad. En este caso lo hace en forma ascendente; si se quiere en orden descendente, se usa (note el signo menos):

```
datos[order(-datos$Age),]
```

Si se requiere ordenar por más de un criterio, digamos edad y altura, se procede como se indica a continuación:

```
datos[order(datos$Age,datos$Height),]
```

A.2.3. Cálculos por niveles de un factor

Si se quiere calcular alguna función (por ejemplo, la media), para cada nivel de un factor (por ejemplo, el sexo), en R se cuenta con varias funciones. Veremos cómo usar la función `by()`.

```
by(datos$NW.Hnd,datos$Sex,mean)
```

calcula la media de la variable `NW.Hnd` para cada nivel de sexo. La orden

```
lapply(datos[,numericas],var)
```

calcula la varianza de cada una de las variables numéricas de datos. La línea de comando

```
sapply(datos[,numericas], var )
```

hace exactamente lo mismo que la anterior, la única diferencia es la forma como R organiza la respuesta. En el primer caso, el objeto regresado por la función es una lista, mientras que en el segundo es un vector de tipo numérico. La siguiente línea de código regresa los mismos valores; en este caso, el número 2 indica que la varianza se va a calcular sobre la segunda dimensión del objeto, es decir, las columnas. Si se usa 1 en lugar de 2, se calcula la media para cada fila. La ventaja de `apply` es que esta función puede emplearse en cálculos con arreglos multidimensionales (`arrays`):

```
apply(datos[,numericas],2,var)
```

Si se quiere una tabla de contingencia cruzando las variables sexo y ejercicio, se puede usar la orden:

```
tapply(datos$Sex,datos[,c("Sex","Exer")],length)
```

Lo anterior es equivalente a

```
table(datos$Sex,datos$Exer)
```

solo que es ventajoso usar `tapply()` porque en el lugar de `length` se puede usar otra función, como `sum` o `var`. Por ejemplo, la edad promedio para cada nivel de sexo y ejercicio se obtiene mediante

```
tapply(datos$Age,datos[,c("Sex","Exer")],mean)
```

La función `agregate()` realiza el mismo trabajo de la función `by()`, solo que la salida de los resultados es un objeto de clase `data.frame`, lo cual es más elegante en cuanto a la presentación.

```
aggregate(datos[,numericas], list(datos$Sex),mean)
```

Con el comando anterior se obtiene la media de las variables numéricas para cada nivel de sexo. El comando siguiente obtiene la mediana de las variables numéricas para cada combinación de niveles de los factores `Sex` y `W.Hnd`:

```
aggregate(datos[,numericas],list(datos$Sex,datos$W.Hnd),
          median)
```

En los dos ejemplos anteriores, las funciones `mean` y `median` se pueden reemplazar por cualquier otra que regrese un valor simple, por ejemplo, `max`, `var`, `length`. Estos son ejemplos de funciones que, en este caso, no se pueden usar con `aggregate`: `cov` y `cor`, ya que el objeto que se entrega es de la clase `data.frame` y el resultado sería una matriz. Esa es una desventaja en relación con la función `by`.

A.3. Cálculo de probabilidades y cuantiles

Cuando se cuenta con un paquete para cálculo estadístico como R, las tablas con algunos cuantiles de las distribuciones no son necesarias. En esta sección y la que sigue, se presenta la forma de calcular probabilidades y percentiles a partir de las distribuciones binomial, de Poisson, normal y ji-cuadrado.

A.3.1. Distribución binomial

Antes de pasar a la distribución binomial, se ilustra cómo realizar un cálculo de combinaciones; la función es `choose()`. Veamos cómo se usa en tres ejemplos:

El valor de	Se obtiene mediante
$\binom{10}{3}$	`choose(10,3)`
$\binom{15}{10}$	`choose(15,10)`
$\binom{100}{4}$	`choose(100,4)`

A.3.1.1. Cálculo de probabilidades a partir de la distribución binomial

Ejemplo: Si X es una variable aleatoria con distribución binomial de parámetros $n = 18$ y $p = 0.1$, calcule $P(X = 2)$.

```
dbinom(x=2,size=18,prob=0.1)
```

Ejemplo: Si X es una variable aleatoria con distribución binomial de parámetros $n = 18$ y $p = 0.1$, calcule $P(X \geq 4) = 1 - P(X \leq 3)$.

```
1-pbinom(3,size=18,p=0.1)
```

Esta probabilidad es equivalente a $P(X > 3)$, la cual se calcula mediante

```
pbinom(3,18,0.1,lower.tail=FALSE)
```

También es equivalente a $\sum_{x=4}^{18} P(X = x)$, que se calcula mediante

```
sum( dbinom(4:18,18,0.1) )
```

Ejemplo: Si X es una variable aleatoria con distribución binomial de parámetros $n = 18$ y $p = 0.1$, calcule $P(3 \leq X < 7)$. Esta probabilidad es igual a $= F(6) - F(2)$; se procede como sigue:

```
pbinom(6,18,0.1)-pbinom(2,18,0.1)
```

Lo anterior es equivalente a

```
sum( dbinom(3:6,18,0.1) )
```

A.3.1.2. Cálculo de cuantiles a partir de la distribución binomial

Si X es una variable aleatoria binomial de parámetros n y p, la función `qbinom()` regresa el valor más pequeño x tal que $P(X \leq x) \geq \alpha$. Veamos un ejemplo:

```
qbinom(0.99,18,0.1)
[1] 5
```

Esto significa que $P(X \leq 5) \geq 0.99$ y 5 es el valor más pequeño que cumple esa desigualdad, con X siendo una variable aleatoria con distribución binomial de parámetros $n = 18$ y $p = 0.1$.

Para la simulación de muestras aleatorias de la distribución binomial se usa la función `rbinom()`.

A.3.2. Distribución de Poisson

Las funciones para la distribución de Poisson son: `dpois()` para probabilidad puntual, `ppois()` para probabilidad acumulada, `qpios()` para cuantiles y `rpois()` para generar muestras aleatorias.

A.3.2.1. Cálculo de probabilidades a partir de la distribución de Poisson

Ejemplo: Si X es una variable aleatoria con distribución de Poisson con media 2.3, calcule $P(X = 2)$.

```
dpois(x=2,lambda=2.3)
```

Ejemplo: Si X es una variable aleatoria con distribución de Poisson con media $E(X) = 4.6$, calcule $P(X \geq 1) = 1 - P(X < 1) = 1 - P(X \leq 0)$.

```
1-ppois(0,4.6)
```

A.3.2.2. Cálculo de cuantiles a partir de la distribución de Poisson

Este caso es similar al de la distribución binomial. Si X es una variable aleatoria Poisson de parámetro λ, la función `qpois()` regresa el valor más pequeño x tal que $P(X \leq x) \geq \alpha$. Por ejemplo:

```
qpois(0.99,2.3)
[1] 6
```

Esto significa que $P(X \leq 6) \geq 0.99$ y 6 es el valor más pequeño que cumple esa desigualdad.

El lector habrá notado el patrón usado por R. Para el cálculo de probabilidad puntual se usa `d` seguido del nombre de la función (`dbinom`, `dpois`); para probabilidad acumulada se usa `p` seguido del nombre de la función; para cuantiles se usa `q`; y para la generación de datos aleatorios se usa `r`. De esa forma es fácil deducir cómo calcular estos valores con las otras distribuciones, por ejemplo, `dgeom()`, `pgeom()`, `qgeom()` y `rgeom()` para la distribución geométrica; `dhyper()`, `phyper()`, `qhyper()` y `rhyper()` para la distribución hipergeométrica.

A.3.3. Distribuciones normal y ji-cuadrado

La sintaxis de R para el cálculo de probabilidades, cuantiles y la generación de números aleatorios, en el caso de las variables aleatorias continuas, es similar al caso discreto. En este, el prefijo `d`, en lugar de calcular probabilidad puntual (que no tiene sentido en el caso continuo), evalúa la función de densidad correspondiente. En este libro se usan frecuentemente las distribuciones normal y ji-cuadrado; por eso se estudian con detalle en las dos secciones siguientes.

A.3.3.1. Distribución normal

Para el cálculo de probabilidades, cuantiles, generación de datos y la evaluación de la función de densidad, a partir de una variable aleatoria X con distribución normal, se cuenta con las funciones `pnorm()`, `qnorm()`, `rnorm()` y `dnorm()`. Veamos cómo usarlas por medio de ejemplos.

Si Z es una variable aleatoria normal estándar, calcule $P(Z \leq 1.65)$.

```
pnorm(1.65)
[1] 0.9505285
```

Note que, por defecto, la función regresa la probabilidad acumulada desde $-\infty$ hasta el valor dado; además, toma media cero y varianza 1 (normal estándar). En general, si la variable aleatoria no es normal estándar, hay que especificar la media y la desviación estándar. Por ejemplo, si X es una variable aleatoria con media 10 y varianza 4, calculemos $P(X < 14)$:

```
pnorm(14,mean=10,sd=2)
[1] 0.9772499
```

Si se desea $P(X > 14)$, se puede hacer de dos formas: $1 - P(X < 14)$ o mediante el comando

```
pnorm(14,mean=10,sd=2,lower.tail=FALSE)
[1] 0.02275013
```

Para el cálculo de cuantiles, veamos el siguiente ejemplo: obtener un valor z tal que si Z es una variable aleatoria normal estándar, se verifique $P(Z < z) = 0.99$:

```
qnorm(0.99)
[1] 2.326348
```

El siguiente código ilustra el uso de la función `dnorm()` para dibujar la función de densidad de una normal estándar.

```
x<-pretty(c(-3.5,3.5),1000) # numeros entre -3.5 y 3.5
y<-dnorm(x) # densidad normal en cada valor de x
plot(x,y,type="l")
```

La siguiente línea de código genera 20 números aleatorios de una distribución normal estándar:

```
dnorm(20)
```

A.3.3.2. Distribución ji-cuadrado

Para el cálculo de la densidad, las probabilidades acumuladas, los cuantiles y la generación de datos de la distribución ji-cuadrado, se cuenta con las funciones `dchisq()`, `pchisq()`, `qchisq()` y `rchisq()`. Veamos cómo usarlas por medio de ejemplos.

Si X^2 es una variable aleatoria ji-cuadrado con un grado de libertad, calcule $P(X^2 > 3.84) = 1 - P(X^2 \leq 3.84)$:

```
1-pchisq(3.84,df=1)
[1] 0.05004352
```

Si X^2 es una variable aleatoria ji-cuadrado con dos grados de libertad, halle el valor x tal que $P(X^2 > x) = 0.05$. Lo que se pide es equivalente a encontrar x tal que $P(X^2 < x) = 0.95$ y se obtiene mediante el comando

```
qchisq(0.95,df=2)
[1] 5.991465
```

Para las demás distribuciones continuas de probabilidad la sintaxis es similar. Para la uniforme, `dunif()`, `punif()`, `qunif()` y `runif()`; para la distribución t de *student*, `dt()`, `pt()`, `qt()` y `rt()`; para la distribución F, `df()`, `pf()`, `qf()` y `rf()`. El usuario solo debe tener en cuenta los parámetros que definen las distribuciones.

Se presentan algunos ejemplos. Si Z es una variable aleatoria normal estándar, entonces

```
pnorm(1.78)  # calcula  \(P(Z\leq 1.78) \)
pnorm(1.78,lower.tail=FALSE) # calcula  \(P(Z \geq 1.78) \)
qnorm(0.972) # calcula \(z\) tal que \(P(Z\leq z)=0.975 \)
dnorm(1.67)  # evalúa la densidad \(\eqref{eq:B.8}\) en \(x=1.67\)
# (\(\mu=0\) y \(\sigma=1\))
rnorm(20)    # genera 20 números aleatorios de la
# distribución normal estándar
```

Si X es una variable aleatoria normal con media $\mu = 10$ y desviación estándar $\sigma = 3$, entonces

```
pnorm(4,mean=10,sd=3)    #calcula  \(P(X\leq 4) \)
qnorm(0.05,mean=10,sd=3) #calcula \(x\) tal que \(P(X\leq x)=0.05\)
dnorm(11,mean=10,sd=3)  #evalúa la densidad \(\eqref{eq:B.8}\) en \(x=11\)
#(\(\mu=10\) y \(\sigma=3\))
rnorm(20,mean=10,sd=3))  #genera 20 números aleatorios de
#la  distribución normal con
#(\(\mu=10\) y \(\sigma=3\))
```

Si X es una variable aleatoria con distribución binomial con parámetros $p = 0.8$ y $n = 20$, entonces

```
pbinom(12,size=20,prob=0.8)   # \(P(X\leq 12)\)
pbinom(12,size=20,prob=0.8,lower.tail=FALSE) # \(P(X\geq 12)\)
qbinom(0.59,size=20,prob=0.8) # valor mas pequeño de \(x\) tal
# que \(P(X\leq x)\geq 0.59\)
dbinom(12,size=20,prob=0.8) # evalúa la función \(\eqref{fdp:binomial}\)
# en \(x=12\) con \(p=0.8\) y \(n=20\)
```

```
# lo que es lo mismo que
# \(P(X=12)\)
rbinom(35,size=20,prob=0.8) # genera una muestra aleatoria
# de  una población con
# distribución binomial de
# parámetros \(p=0.8\) y \(n=20\)
```

Este esquema se repite con las demás distribuciones. La tabla A.1 muestra la sintaxis de las funciones para las distribuciones de uso frecuente en análisis de sobrevida.

Tabla A.1. Distribuciones de probabilidad con R

Función Densidad	Func Prob. Acumulada	Cuantiles	Números Aleatorios
Uniforme			
dunif(q,min,max)	punif(q, a, b)	qunif(p,a,b)	runif(N,a,b)
Normal			
dnorm(q,min,max)	pnorm(q, μ, σ)	qnorm(p, μ, σ)	rnorm(N,μ, σ)
Binomial			
dbinom(q,n,pi)	punif(q, n, π)	qbinom(p,n,π)	rbinom(N,n,π)
log-normal			
dlnorm(q, μ, σ)	pulnorm(q, μ, σ)	qlnorm(p,μ, σ)	rlnorm(N,μ, σ)
Beta			
dbeta(q,α, β)	pbeta(q, α, β)	qbeta(p,α, β)	rbeta(N,α, β)
Geométrica			
dgeom(q,π)	pgeom(q, π)	qgeom(p,π)	rgeom(N,π)
Gama			
dgamma(q,α, β)	pgamma(q, α, β)	qgamma(p,α, β)	rgamma(N,α, β)
ji-cuadrado			
dchisq(q,gl)	pchisq(q,gl)	qchisq(p,gl)	rchisq(N,gl)
Exponencial			
dexp(q,θ)	dexp(q,θ)	qexp(p,θ)	rexp(N,θ)
F			
df(q,gl1,gl2)	pf(q,gl1,gl2)	qf(p,gl1,gl2)	rf(N,gl1,gl2)
Hipergeométrica			
dhyper(q,m,n,k)	phyper(q,m,n,k)	qhyper(p,m,n,k)	rhyper(N,m,n,k)
t-Student			
dt(q,gl)	pt(q, gl)	qt(p, gl)	rt(N,gl)
Poisson			
dpois(q,λ)	ppois(q,λ)	qpois(p,λ)	rpois(N,λ)
Weibull			
dweibull(q,θ, α)	pweibull(q,θ, α)	qweibull(p,θ, α)	rweibull(N,θ, α)

Referencias

[1] O. O. Aalen, *Nonparametric inference for family of counting processes*, Annals of Statistics **6** (1978), 534–545.

[2] A. Agresti, *Categorical data analysis*, John Wiley, 2019.

[3] G. A. Bohoris, *Comparison of the cumulative-hazard and kaplan-meier estimators of survival function*, IEEE Transactions on Reliability **43** (1994), no. 1, 230–232.

[4] G. E. P Box and G. C. Tiao, *Bayesian inference in statistical analysis*, John Wiley and Sons, 1973.

[5] N. E. Breslow, *Covariance analysis of censored survival data*, Biometrics **30** (1974), 89–100.

[6] E. R. Brown and J. G. Ibrahim, *A bayesian semiparametric joint hierarchical model for longitudinal and survival data*, Biometrics **59** (2003), 221–228.

[7] T. Burzykowski, G. Molenberghs, and M. Buyse, *The evaluation of surrogate endpoints*, Springer, 2005.

[8] J. Cai and D. E. Schaubel, *Analysis of recurrent events data*, Elsevier B. V. Handbook of Statistics **23** (2004), 603–623.

[9] C. Chatfield, *Statistics for Technology. A Course in Applied Statistics*, Springer-Science+Business Media, B.V., 1978.

[10] D. Collet, *Modelling survival data in medical research*, Chapman and Hall/CRC, 2014.

[11] E. A. Colosimo and S. R. Giolo, *Análise de sobrevivencia aplicada*, Edgard Blucher, 2006.

[12] P. Congdon, *New York: John Wiley & Sons*, Bayesian Statistical Modelling, 2001.

[13] R. J. Cook and J. F Lawless, *The statistical analysis of recurrent events*, Springer, 2007.

[14] D. R. Cox, *Regression models a life tables*, Journal of the Royal Statistical Society, Ser. B **34** (1972), 187–220.

[15] _____, *Partial likelihood*, Biometrika **62** (1975), 269–276.

[16] D. R Cox and D. Oakes, *Analysis of survival data*, Chapman and Hall/CRC, 1994.

[17] M. Crowder, *Multivariate survival analysis and competing risks*, 2a. ed., Chapman and Hall/CRC, 2012.

[18] V. DeGruttola and X. Tu, *Modelling the progression of CD4-lymphocyte count and its relationship to survival time*, Biometrics **50** (1994), 1003–1014.

[19] A. Dempster, N. Lair, and D. Rudin, *Maximum likelihood from incomplete data via the em-algorithm*, Statistica Sinica **39** (1977), 1–38.

[20] L. G. Díaz and L. R. León, *Análisis estadístico de medidas repetidas*, Facultad de Ciencias - Universidad Nacional de Colombia, 2018.

[21] L. G. Díaz and M. A. Morales, *Análisis estadístico de datos multivariados*, Facultad de Ciencias - Universidad Nacional de Colombia, 2015.

[22] L. G. Díaz, M. A. Morales, and L-R. Leon, *Análisis estadístico de datos categóricos*, Universidad Nacional de Colombia, 2018.

[23] P. J. Diggle, P. Heagerty, K. Y. Liang, and S. Zeger, *Analysis of longitudinal data*, Oxford University Press, 2002.

[24] R. M. Elashoff, G. Li, and N. Li, *Joint modeling of longitudinal and time to event data*, Chapman and Hall, 2017.

[25] C. L. Faucett and D. C. Thomas, *Simultaneously modeling censored survival data and repeatedly measured covariates: a gibbs sampling approach*, Statistics in Medicine **15** (1996), 1663–1685.

[26] G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, *Longitudinal data analysis*, Chapman and Hall/CRC, 2009.

[27] M. H. Gail, T. J. Santner, and C. C. Brown, *An analysis of comparative carcinogenesis experiments based on multiple times to tumor*, Biometrics **36** (1980), 255–266.

[28] D. Gamerman, *Markov chain monte carlo*, Chapman and Hall, 2006.

[29] E. J. Gumbel, *Bivariate exponential distributions*, J. Am. Statis. Assoc. **55** (1960), 698–707.

[30] D. D. Hanagal, *Modeling survival data using frailty models*, Chapman and Hall/CRC, 2011.

[31] R. Henderson, P. Diggle, and A. Dobson, *Joint modelling of longitudinal measurements and event time data*, Biostatistics **4** (2000), 465–480.

[32] R. V. Hogg, J. W. McKean, and A. T. Craig, *Introduction to mathematical statistics*, Pearson Prentice Hall, 2005.

[33] P. Hougaard, *Modeling survival data using frailty models*, Chapman and Hall/CRC, 2000.

[34] J. G. Ibrahim, M. H. Chen, and D. Sinha, *Bayesian survival analysis*, Springer, 2010.

[35] H. Joe, *Dependence modeling with copulas*, Chapman and Hall/CRC, 2015.

[36] M. Kaeding, *Bayesian analysis of failure time data using p-splines*, Springer Spektrum, 2015.

[37] J. D. Kalbfleisch and R. L. Prentice, *The statistical analysis of failure time data*, John Wiley and Sons, 2011.

[38] J. P. Klein, *Small-sample moments of some estimators of the variance of the kaplan-meier and nelson-aalen estimators*, Scandinavian Journal of Statistics **18** (1991), 333–340.

[39] J. P. Klein and M. L. Moeschberger, *Survival analysis. techniques for censured and truncated data*, Springer, 2005.

[40] C. D Lai, *Generalized weibull distributions*, Springer, 2014.

[41] N. M. Laird and J. H. Ware, *Random effects models for longitudinal data*, Biometrics **38** (1982), 963–974.

[42] T Lancaster, *Econometric methods for the duration of unemployment*, Econometrica **47** (1979), 939–956.

[43] J. F. Lawless, *Statistical models and methods for lifetime data*, John Wiley and Sons, 2003.

[44] E. T. Lee and J. W. Wang, *Statistical methods for survival data analysis*, John Wiley and Sons, 2003.

[45] V. Leyva, *Birnbaum saunders distribution*, Academic Press, 2016.

[46] K. Y. Liang and S. L. Zeger, *Longitudinal data analysis using generalized linear models*, Biometrika **73** (1986), 13–22.

[47] D. Y. Lin and Z. Ying, *Semiparametric analysis of the additive risk model*, Statistics in Medicine **81** (1994), 61–71.

[48] _____, *Semiparametric Analysis of general additive-multiplicative hazard models for counting processes*, Annals Statistics **23** (1995), 1712–1734.

[49] _____, *Additive regression models for survival data*, In Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis, D. Y. Lin and T. R. Fleming **1** (1997), 185–198.

[50] L. Liu, R. A. Wolfe, and X. Huang, *Shared frailty models for recurrent events and a terminal event*, Biometrics **60** (2004), 747–756.

[51] X. Liu, *Methods and applications of longitudinal data analysis*, Academic Press, 2015.

[52] A. W. Marshall and I. Olkin, *Families of multivariate distributions*, J. Am. Statis. Assoc. **83** (1988), 834–841.

[53] P. McCullagh and J. A. Nelder, *Generalized linear models*, Champman & Hall, 1989.

[54] C. McCulloch and S. Searle, *The em algorithm and extensions*, John Wiley and Sons, 2001.

[55] D. F. Moore, *Applied survival analysis using r*, Springer, 2016.

[56] R. B. Nelsen, *An introduction to copulas*, Springer, 2006.

[57] W. Nelson, *Theory and applications of hazard plotting for censuring failure data*, Technometrics **14** (1972), no. 1, 945–965.

[58] G. G. Nielsen, R. D. Gill, P. K. Andersen, and T. I. A. Sorensen, *A counting process approach to maximum likelihood estimation in frailty models*, Scandinavian Lournal of Statistics **19** (1992), 25–43.

[59] R. Prentice, *Surrogate endpoints in clinical trials: definition and operation criteria*, Statistics in Medicine **8** (1989), 431–440.

[60] S. J. Ratcliffe, G. Wensheng, and T. H. Thomas R., *Joint modeling of longitudinal and survival data via a common frailty*, Biometrics **60** (2004), no. 4, 892–899.

[61] D. Rizopoulos, *Joint modeling of longitudinal and time to event data: With applications in r*, Chapman and Hall, 2012.

[62] SAS/STAT, *User's guide the lifereg procedure*, SAS, 2015.

[63] D. A. Schoenfeld, *Partial residuals for the proportional hazard regressiom model*, Biometrika **69** (1982), 239–241.

[64] P. J. Smith, *Analysis of failure and survival data*, Chapman and Hall/CRC, 2017.

[65] M. E. Stokes, C. S. Davis, and G. G. Koch, *Categorical data analysis using the sas system*, Springer, 2003.

[66] T. M. Therneau and P. M. Grambsch, *Modeling Survival Data. Extending the Cox Model*, Springer, 2000.

[67] A. Tsiatis and M Davidian, *Joint modeling of longitudinal and time to event data: an overview*, Journal of the Royal Statistical Society **14** (2004), 809–834.

[68] G. Tutz and M Schmid, *Modeling discrete time-to-event data*, Springer, 2016.

[69] J. W. Vaupel, K. G Manton, and E. Stallard, *The impact of heterogeneity in individual frailty on the mortality*, Demographiy **16** (1979), 439–454.

[70] G. Verbeke and G. Molenberghs, *Linear mixed models for longitudinal data*, Springer, 2000.

[71] ———, *Models for discrete longitudinal data*, Springer, 2005.

[72] A. Wienke, *Frailty Models in Survival Analysis*, Chapman and hall/CRC, 2011.

[73] M. S. Wulfsohn and A. A. Tsiatis, *A joint model for survival and longitudinal data measured with error*, Biometrics **53** (1997), 330–339.

Colección
Textos

Este texto va dirigido para un lector que quiera conocer y entender los conceptos del análisis estadístico de datos de sobrevida, como también el manejo y el uso de herramientas estadísticas, útiles para la exploración, descripción y el análisis confirmatorio de datos, principalmente del tiempo hasta la ocurrencia de un evento. El material estadístico ofrecido en este texto puede ser empleado por investigadores de varias disciplinas, pues basta cambiar el escenario de los ejemplos e ilustraciones, para hacer de este texto un instrumento de apoyo para áreas de afines tales como el análisis de confiabilidad, análisis del riesgo, el análisis de degradación, durabilidad o longevidad y el control estadístico de calidad. El rigor matemático usado en el texto es el básico; solo se requiere de los conocimientos contenidos en un curso de probabilidad e inferencia estadística. Se enuncian los conceptos, definiciones y expresiones probabilísticas asociadas con los datos de sobrevida. La estimación se presenta desde perspectiva paramétrica, no paramétrica y bayesiana. Se desarrolla el modelamiento estadístico, desde una visión tanto paramétrica, semi-paramétrica como bayesiana. Se incluye una introducción al análisis de sobrevida de datos multivariados. Se emplea el paquete R para el desarrollo computacional de las diferentes técnicas estadísticas asociadas.

Estadística