Universidad de Granada

Departamento de Estadística e Investigación Operativa

Tesis Doctoral

Programa de doctorado en Matemáticas y Estadística

Aportaciones en el Análisis de Fiabilidad mediante técnicas no paramétricas

Antonio Jesús López Montoya

Granada, 2017

Editor: Universidad de Granada. Tesis Doctorales
Autor: Antonio Jesús López Montoya
ISBN: 978-84-9163-225-2
URI: http://hdl.handle.net/10481/46755

Tesis Doctoral

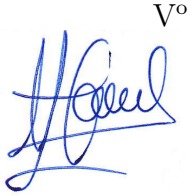
Aportaciones en el Análisis de Fiabilidad mediante técnicas no paramétricas

Memoria de tesis presentada por D. Antonio Jesús López Montoya para optar al grado de Doctor por la Universidad de Granada.

Esta Tesis Doctoral ha sido dirigida por las Doctoras Dña. María Luz Gámiz Pérez, y Dña. María Dolores Martínez Miranda, Profesoras Titulares de Universidad del Departamento de Estadística e Investigación Operativa de la Universidad de Granada.

Granada, 22 de febrero de 2017

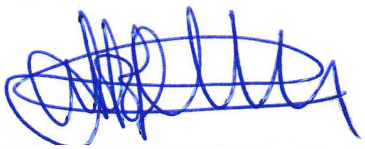
Vº. Bº. de las directoras de la Tesis



Fdo.: María Luz Gámiz Pérez

Fdo.: María Dolores Martínez Miranda

Doctorando



Fdo.: Antonio Jesús López Montoya

El doctorando Antonio Jesús López Montoya y las directoras de la tesis María Luz Gámiz Pérez y María Dolores Martínez Miranda garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección de las directoras de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

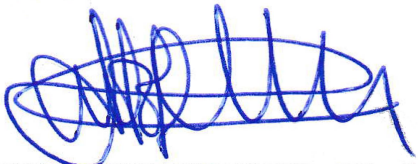
Granada, 22 de febrero de 2017

Directoras de la Tesis

Fdo.: María Luz Gámiz Pérez

Fdo.: María Dolores Martínez Miranda

Doctorando



Fdo.: Antonio Jesús López Montoya

Agradecimientos

En primer lugar, quiero expresar mis agradecimientos a mis directoras de tesis, María Luz Gámiz Pérez y María Dolores Martínez Miranda, por haberme brindado la oportunidad de realizar este trabajo, así como el gran esfuerzo, dedicación, apoyo y sobre todo la infinita paciencia que me han dedicado para que este trabajo se llevase a cabo. A la profesora Rocío Raya Miranda por su inestimable ayuda en la realización del Capítulo 3 de este trabajo. Al Departamento de Estadística e Investigación Operativa de la Universidad de Granada por darme la oportunidad de realizar este trabajo. I would also like to thank Dr. Jens Perch Nielsen for the guidance he gave us during his stay in Granada in 2012 and whose works have inspired and motivated a large part of this work.

A mi querida Irene por estar siempre ahí "al pie del cañón" ofreciéndome siempre su incondicional apoyo, comprensión, paciencia, ánimo y un largo etcétera para el cual no hay páginas sufiecientes en este trabajo. También a sus padres José Carlos y Eva María así como a su hermana Laura que siempre se preocupan por nosotros y nos ayudan en todo y porque la expresión "buenas personas" se les queda corta. A su hermano Álvaro, por su gran ayuda en el diseño y realización de la portada de este trabajo.

A mi padre, que siempre hizo lo imposible para inculcarnos la afición por el conocimiento, a mi madre por secundarlo, a mi hermana Mari porque sus palabras siempre sirven de esperanza en los malos momentos, a mi hermana Toñi porque siempre me ha ayudado en todo así como sus sabios consejos, a mi hermana Ani

por su incombustible ayuda y disposición, a mi hermana Carmen porque me ha acompañado siempre, y a mi tío Fernando por ser nuestro segundo padre.

A mis amigos, Kike, Rocío y Juanico por estar siempre ahí cuando los necesito, a mi primo Miguel por los buenos momentos que siempre pasamos juntos. También nombrar a Rosita y Fran, Lydia y Antonio, Elvira, Fernando, Isa, María, Javi y Concha porque con ellos siempre he podido evadirme de la rutina. A mi amigo Antonio Caparrós por ayudarme a digitalizar bases de datos.

A mis compañeros de trabajo de las Salvamares Algenib, Denébola y Sirius por lo bien que me han hecho sentir siempre en mi trabajo y los grandes e inolvidables momentos que hemos vivido tanto profesionales como personales durante estos años.

Finalmente, pero no por ello menos importante, dar las gracias a mi cuñado Manolo ya que sin su ayuda, quizás, este trabajo no hubiese sido posible.

Índice general

Resumen **1**

Notación y definiciones **4**

1. Conceptos básicos **9**

1.1. Introducción 9

1.2. Herramientas probabilísticas 10

1.2.1. Conceptos básicos 11

1.2.2. Algunas nociones sobre teoría de martingalas 15

1.3. El modelo multiplicativo de Aalen 19

1.3.1. Procesos de recuento con un único salto 20

1.3.2. Procesos de recuento para la modelización de eventos recurrentes 26

1.4. El proceso de Poisson no homogéneo (PPNH) 28

1.5. Inferencia en procesos de recuento 32

1.5.1. Enfoque paramétrico 32

1.5.2. Enfoque no paramétrico 34

1.6. Estimación tipo núcleo 37

1.6.1. Estimador tipo núcleo y criterios de error 37

1.6.2. Estimación de la función de intensidad de un proceso de recuento 41

1.6.3. Estimación de la función de supervivencia 46

1.6.4. Estimadores locales mínimo-cuadráticos 48
1.6.5. El parámetro de suavizado en la estimación de la función de intensidad 55

2. Modelos de regresión con datos de supervivencia basados en procesos de recuento 61

2.1. Introducción y objetivos 61
2.2. Modelos de regresión basados en la función de azar 64
2.2.1. Modelo de riesgos proporcionales (PH) 68
2.2.2. Modelo de riesgos aditivos de Aalen 73
2.2.3. Modelo de tiempo de vida acelerada (AFT) 75
2.2.4. Regresión no paramétrica de la función de azar 76
2.3. Estimación no paramétrica del modelo AFT 80
2.3.1. Estimación de los coeficientes de regresión 82
2.3.2. Estimación del error estándar 84
2.3.3. Estimación no paramétrica de la función de supervivencia base 85
2.4. Estudio de simulación 92
2.4.1. Objetivos y especificaciones del estudio 92
2.4.2. Generación de datos de supervivencia 94
2.4.3. Criterios de error 94
2.4.4. Resultados y conclusiones 95
2.5. Aplicación con datos reales 101
2.5.1. Presentación de los datos y objetivos del análisis 101
2.5.2. Verificación de la hipótesis de riesgos proporcionales 103
2.5.3. Ajuste del modelo AFT 106
2.6. Conclusiones 110

3. Inferencia en un espacio de localización y escala de intensidades en PPNH 111

3.1. Introducción y objetivos 111
3.2. Ejemplos de datos reales modelizados mediante un PPNH 114
3.3. Estimador lineal local de la función de intensidad de un PPNH . . . 116
3.4. Inferencia en un espacio de localización y escala 119
3.4.1. La herramienta exploratoria SiZer Map de Chaudhuri y Marron (1999) 119
3.4.2. SiZer Map para la función de intensidad de un PPNH 121
3.5. Estudio de Simulación 126
3.5.1. Descripción de los modelos simulados 126
3.5.2. Evaluación de SiZer para la detección de cambios en la tendencia 133
3.5.3. Cobertura de los intervalos de confianza 134
3.5.4. Resultados y conclusiones 137
3.6. Aplicaciones con datos reales 145
3.6.1. Datos de ocurrencia de réplicas en terremotos 145
3.6.2. Datos de un sistema hidráulico de máquinas de carga, acarreo y descarga 149
3.6.3. Datos del sistema del tren de potencia de un autobús urbano 150
3.6.4. Datos de temporales sucedidos en el mar Ártico 151
3.6.5. Datos de inmigración ilegal 153
3.6.6. Datos de accidentes en minas de carbón 155
3.6.7. Un análisis diagnóstico de la hipótesis de PPNH en los datos reales 157
3.7. Resultados y conclusiones 162

Bibliografía **164**

Apéndices **175**

A. Código en R del estudio de simulación del Capítulo 2 **175**

B. Código en R del estudio de simulación del Capítulo 3 **203**

Índice de figuras

1.1. Representación gráfica de datos de tiempos de vida: completos, censurados por la derecha, truncados por la izquierda y combinación de datos censurados por la derecha y truncados por la izquierda. 22

1.2. Representación gráfica de los procesos $N(t)$ e $Y(t)$, para datos completos, censura por la derecha y truncamiento por la izquierda. . . . 26

1.3. Datos de eventos recurrentes: tiempos de fallo (horas) del sistema hidráulico de seis maquinas de LHD. 27

1.4. Funciones de intensidad estimadas con parámetros de suavizado obtenidos por el método de validación cruzada y por el método plug-in. Datos de fallo de software. 59

2.1. Caso 1. Modelo Exponencial. Box-plots de las estimaciones del ASE para los modelos PH y AFT con $n = 50$, 100 y 200, con nivel de truncamiento $p_t = 10\,\%$ y niveles de censura $p_c = 10\,\%$, $30\,\%$ y $45\,\%$. 98

2.2. Caso 2. Modelo Weibull con parámetro de forma igual a 5. Box-plots de las estimaciones del ASE para los modelos PH y AFT con $n = 50$, 100 y 200, con nivel de truncamiento $p_t = 10\,\%$ y niveles de censura $p_c = 10\,\%$, $30\,\%$ y $45\,\%$. 99

2.3. Caso 3. Modelo Weibull con parámetro de forma igual a 0.5. Boxplots de las estimaciones del ASE para los modelos PH y AFT con $n = 50$, 100 y 200, con nivel de truncamiento $p_t = 10\,\%$ y niveles de censura $p_c = 10\,\%$, $30\,\%$ y $45\,\%$. 100

2.4. Funciones de azar acumulado estratificadas por las covariables material, tráfico, longitud y diámetro. 104

2.5. Gráfico de los residuos escalados de Schoenfeld para las covariables material, tráfico, longitud y diámetro. 105

2.6. Gráfico de los residuos de martingala para las covariables longitud y diámetro. 107

2.7. Estimador lineal local de la función de supervivencia base. 109

3.1. Análisis SiZer con intervalos de confianza bootstrap de tipo simultáneo para datos de un sistema hidráulico. 122

3.2. Representación gráfica de los cuatro modelos simulados: De izquierda a derecha, la primera fila muestra los modelos M1 y M2, la segunda fila muestra los modelos M3 y M4. 127

3.3. Análisis SiZer con intervalos de confianza bootstrap de tipo simultáneo para datos simulados del modelo M1. 129

3.4. Análisis SiZer con intervalos de confianza bootstrap de tipo puntual para datos simulados del modelo M2. 130

3.5. Análisis SiZer con intervalos de confianza bootstrap de tipo puntual para datos simulados del modelo M3. 131

3.6. Análisis SiZer con intervalos de confianza Normal de tipo puntual para datos simulados del modelo M4. 132

3.7. Cobertura empírica puntual. a) Cobertura de los Intervalos de confianza de tipo puntual construidos mediante la aproximación Normal. b) Cobertura de los Intervalos de confianza de tipo puntual construidos mediante la aproximación bootstrap. Resultados con 1000 muestras de tamaño $n = 100$ del modelo M1. 139

3.8. Cobertura empírica global. S.B. denota intervalos bootstrap de tipo simultáneo. S.N. intervalos de tipo simultáneo construidos mediante la aproximación Normal. P.N. intervalos de tipo puntual construidos mediante la aproximación Normal. P.B. intervalos bootstrap de tipo puntual. Muestras de tamaño $n = 100$ del modelo M1. 140

3.9. Cobertura empírica puntual. a) Cobertura de los Intervalos de confianza de tipo puntual construidos mediante la aproximación Normal. b) Cobertura de los Intervalos de confianza de tipo puntual construidos mediante la aproximación bootstrap. Resultados con 1000 muestras de tamaño $n = 500$ del modelo M2. 141

3.10. Cobertura empírica global. S.B. denota intervalos bootstrap de tipo simultáneo. S.N. intervalos de tipo simultáneo construidos mediante la aproximación Normal. P.N. intervalos de tipo puntual construidos mediante la aproximación Normal. P.B. intervalos bootstrap de tipo puntual. Muestras de tamaño $n = 500$ del modelo M2. 142

3.11. Cobertura empírica puntual. a) Cobertura de los Intervalos de confianza de tipo puntual construidos mediante la aproximación Normal. b) Cobertura de los Intervalos de confianza de tipo puntual construidos mediante la aproximación bootstrap. Resultados con 1000 muestras de tamaño $n = 500$ del modelo M3. 143

3.12. Cobertura empírica puntual. a) Cobertura de los Intervalos de confianza de tipo puntual construidos mediante la aproximación Normal. b) Cobertura de los Intervalos de confianza de tipo puntual construidos mediante la aproximación bootstrap. Resultados con 1000 muestras de tamaño $n = 1000$ del modelo M4. 144

3.13. Análisis SiZer con intervalos de confianza Normal de tipo simultáneo para datos de réplicas del terremoto de magnitud 6.2 en la escala de Richter sucedido el 26 de julio de 2003 en el norte de Miyagi-Ken (Japón). 147

3.14. Análisis SiZer con intervalos de confianza Normales de tipo simultáneo para datos de réplicas del terremoto de magnitud 7.9 en la escala de Richter sucedido el 12 de mayo de 2008 en Sichuan (China). 148

3.15. Análisis SiZer con intervalos de confianza bootstrap de tipo simultáneo para datos de un sistema hidráulico. 150

3.16. Análisis SiZer con intervalos de confianza bootstrap de tipo simultáneo para datos de un sistema de tren de potencia de un autobús urbano. 152

3.17. Análisis SiZer con intervalos de confianza Normal de tipo puntual para datos relativos a las tormentas sucedidas en el mar Ártico. . . 154

3.18. Análisis SiZer con intervalos de confianza Normal de tipo simultáneo para datos relativos al rescate de pateras en aguas españolas. 156

3.19. Análisis SiZer con intervalos de confianza bootstrap de tipo simultáneo para datos relativos al rescate de pateras en aguas españolas, realizados separadamente por años, siendo los años (a)= 2012, (b)= 2013, (c)= 2014 y (d)= 2015. 159

3.20. Análisis SiZer con intervalos de confianza Normal de tipo puntual para datos relativos a los accidentes ocurridos en minas de carbón. . 160

3.21. Diagnóstico gráfico de la hipótesis de PPNH para los conjuntos de datos analizados siendo (a) terremotos en Japón, (b) terremotos en China, (c) sistema hidráulico, (d) tren de potencia de un autobus urbano, (e) rescate de pateras y (f) accidentes en minas de carbón. . 161

Índice de tablas

2.1. Aproximación de Monte Carlo del MSE para los modelos PH, AFT y Exponencial. 96
2.2. Aproximación de Monte Carlo del MSE para los modelos PH, AFT y Weibull(1, 5). 96
2.3. Aproximación de Monte Carlo del MSE para los modelos PH, AFT y Weibull(1, 0.5). 96
2.4. Resultados del test sobre la hipótesis de riesgos proporcionales. . . . 103
2.5. Estimación de los coeficientes de regresión del modelo semi-paramétrico AFT. 107

3.1. Descripción de los modelos teóricos. 126
3.2. Recuentos observados en 1000 muestras de tamaño $n = 100$ simuladas a partir del modelo M1. Se muestran por columnas los niveles de suavizado considerados en escala logarítmica. Las filas en color gris muestran el número de veces que SiZer detecta el número real de cambios en la tendencia de la función de intensidad. 135
3.3. Recuentos observados en 1000 muestras de tamaño $n = 500$ simuladas a partir del modelo M2. Se muestran por columnas los niveles de suavizado considerados en escala logarítmica. Las filas en color gris muestran el número de veces que SiZer detecta el número real de cambios en la tendencia de la función de intensidad. 136

3.4. Recuentos observados en 1000 muestras de tamaño $n = 500$ simuladas a partir del modelo M3. Se muestran por columnas los niveles de suavizado considerados en escala logarítmica. Las filas en color gris muestran el número de veces que SiZer detecta el número real de cambios en la tendencia de la función de intensidad. 137
3.5. Recuentos observados en 1000 muestras de tamaño $n = 1000$ simuladas a partir del modelo M4. Se muestran por columnas los niveles de suavizado considerados en escala logarítmica. Las filas en color gris muestran el número de veces que SiZer detecta el número real de cambios en la tendencia de la función de intensidad. 138

Resumen

Esta tesis doctoral se centra en el uso de modernos métodos estadísticos para evaluar las características de funcionamiento de un sistema de fiabilidad dinámico que opera bajo ciertas condiciones y que evoluciona aleatoriamente en el tiempo.

En concreto, se proponen nuevos modelos estocásticos que permiten relacionar las medidas de funcionamiento de un sistema con el entorno físico en el que éste trabaja. Además de la explicación de la formulación matemática subyacente en estos modelos, se acompaña con un análisis de numerosos ejemplos ilustrativos a partir de diferentes estudios de simulación así como de análisis de datos reales.

Los métodos considerados en esta tesis se enmarcan dentro de la estadística no paramétrica, que no es el enfoque habitual cuando se tratan problemas de inferencia en el contexto del Análisis de Fiabilidad. Tradicionalmente el Análisis de Fiabilidad se ha basado en métodos estadísticos en los que se asumen determinadas hipótesis (paramétricas) sobre la estructura que subyace en los datos a analizar. Sin embargo, hay muchas situaciones prácticas en las que esas hipótesis no se pueden mantener y de hecho, muchas veces un enfoque no paramétrico del problema es la única opción razonable.

En este sentido, esta tesis propone dos nuevas contribuciones al Análisis de Fiabilidad desde la perspectiva de los métodos de Estadística no paramétricos.

La memoria está estructurada en tres capítulos que se resumen a continuación.

En el Capítulo 1 se presentan conceptos básicos utilizados en los demás capítulos. Se considera una formulación basada en procesos de recuento para modelizar

los procesos de tiempos de fallo que se analizarán en esta memoria. Usando esta formulación se construyen los estimadores de las características de interés relativas al tiempo de fallo. Los datos que usualmente se registran en Análisis de Fiabilidad y Supervivencia presentan características especiales como pueden ser censura y/o truncamiento y que suponen que la información contenida en la muestra es en cierto modo incompleta con respecto al fenómeno real que se pretende estudiar. Bajo esta perspectiva de procesos de recuento, se recogen en este capítulo una serie de aportaciones recientemente propuestas en la literatura en las que se formulan estimadores no paramétricos tipo núcleo para estudiar las características del tiempo de vida más interesantes para el análisis de la fiabilidad de un sistema, en particular la función de azar y la función de fiabilidad o supervivencia.

El Capítulo 2 presenta un modelo de regresión de tiempo de vida acelerada (AFT) no paramétrico que permite evaluar el efecto de determinados factores sobre la probabilidad de que un sistema sobreviva un determinado periodo de tiempo. La motivación práctica para este estudio es el análisis del rendimiento de una red de suministro de agua situada en una ciudad de la costa mediterránea española, para lo cual se dispone de un conjunto de datos relativos a los tiempos de rotura de tramos de tubería de la citada red.

A diferencia de la metodología comúnmente utilizada que impone restrictivas hipótesis de tipo paramétrico, en este capítulo se propone un modelo no paramétrico que no especifica ninguna distribución en particular para los tiempos de fallo subyacentes. Para llevar a cabo el ajuste del modelo, se sugiere un procedimiento que consta de dos etapas. En la primera se analiza la incidencia de ciertos factores sobre el tiempo de vida del sistema. En la segunda, se introduce un estimador no paramétrico de la función de supervivencia base a través de la formulación de los datos basada en procesos de recuento. Se obtienen las propiedades asintóticas de este estimador y se evalúa el comportamiento en muestras finitas a través de un extenso estudio de simulación que muestra que la metodología propuesta ofrece

mejores resultados que otras propuestas semi-paramétricas que habitualmente se utilizan en la práctica.

El trabajo desarrollado en este capítulo ha sido publicado por la revista Applied Mathematical Modelling (López-Montoya, Gámiz-Pérez y Martínez-Miranda (2015)). Los resultados obtenidos en este capítulo han sido presentados en el *27th International Workshop on Statistical Modelling (IWSM)* celebrado en Praga, República Checa y en el *XXXIII Congreso Nacional de Estadística e Investigación Operativa y VII Jornadas de Estadística Pública (SEIO)* celebrado en Madrid, España.

En el Capítulo 3 se presenta una extensión de la herramienta exploratoria gráfica SiZer Map para hacer inferencia sobre la función intensidad de un proceso de Poisson no homogéneo (NHPP). En este caso, la motivación práctica es el análisis de la razón de ocurrencia de fallos (ROCOF) de un sistema de fiabilidad con reparación mínima. La extensión de SiZer se define considerando un estimador tipo núcleo lineal local para la función intensidad así como para su derivada. El parámetro de suavizado es la escala de visualización y el intervalo de tiempo donde se realiza la estimación es el espacio de localización. En este capítulo se ilustra el uso de dicha herramienta gráfica con varios conjuntos de datos de situaciones reales. Además se presenta un estudio de simulación que revela que SiZer es una herramienta adecuada para la detección de cambios significativos en la tendencia de la función intensidad.

El trabajo desarrollado en este capítulo ha sido sometido y está en proceso de revisión (Gámiz-Pérez, López-Montoya, Martínez-Miranda y Raya-Miranda (2017)). Los resultados obtenidos en este capítulo han sido presentados en el *The 3rd Conference of the International Society for Non-Parametric Statistics (ISNPS)* celebrado en Avignon, Francia y en el *XXXV Congreso Nacional de Estadística e Investigación Operativa y IX Jornadas de Estadística Pública (SEIO)* celebrado en Pamplona, España.

Notación y definiciones

- Ω un espacio arbitrario, normalmente con las salidas de un experimento.
- $\mathcal{F}$ una σ-álgebra de conjuntos de Ω, llamada filtración.
- $\mathcal{F}_t \equiv \sigma\{X(s) : 0 \leq s \leq t\}$ es la menor σ-álgebra tal que $X(s)$ es medible para cada $s \in [0, t]$, y $\{\mathcal{F}_t : t \geq 0\}$ es la historia del proceso estocástico X.
- $\mathcal{F}_{t-}$ la menor σ-álgebra que contiene a todos los conjuntos en $\bigcup_{a>0} \mathcal{F}_{t-a}$, también escrito de la forma $\sigma\left\{\bigcup_{a>0} \mathcal{F}_{t-a}\right\}$ o también $\bigvee_{a>0} \mathcal{F}_{t-a}$.
- P medida de probabilidad en Ω.
- $\mathcal{B}$ σ-álgebra de Borel en el dominio de los números reales, esto es, la menor σ-álgebra que contiene a todos los intervalos abiertos de $\mathbb{R}$.
- $Pr\{a|b\}$ probabilidad de a condicionada a b.
- *càdlàg* proceso estocástico cuyas trayectorias son continuas a la derecha con límites a la izquierda.
- $D([0, t])$ espacio de Skorokhod que consiste en funciones de tipo *càdlàg* en el intervalo $[0, t]$.
- $\mathbf{A}$ indica un vector o una matriz.
- $\mathbf{A}^\top$ traspuesta de la matriz $\mathbf{A}$.

- diag($\mathbf{A}$) matriz diagonal de $\mathbf{A}$.
- $I[B]$ función que asigna el valor 1 si se cumple la condición B y 0 si no se cumple.
- $N(t)$ proceso de recuento.
- $Y(t)$ proceso de riesgo.
- $\lambda(t)$ función de intensidad.
- $\alpha(t)$ función de azar o de riesgo.
- v.a. variable aleatoria.
- i.i.d. independientes e idénticamente distribuidos.
- h parámetro de suavizado.
- K función núcleo.
- $\mu_j(K) = \int_{-\infty}^{\infty} u^j K(u)du$ momento de orden j del núcleo K.
- $R(K) = \int_{-\infty}^{\infty} K^2(u)du$ función de curvatura del núcleo K.
- $m^{(\nu)}$ derivada ν-ésima de la función m.
- $C^{(\nu)}([0,t])$ espacio de funciones con ν derivadas continuas en $[0,t]$.
- $\triangleq$ igual por definición.
- $\equiv$ equivalente.
- $\approx$ aproximadamente igual.
- $\propto$ directamente proporcional.
- CV validación cruzada.

- Bias sesgo.
- ASE error cuadrático promedio.
- ISE error cuadrático integrado.
- MSE error cuadrático medio.
- MISE error cuadrático medio integrado.
- AMISE error cuadrático medio integrado asintótico.
- LTRC truncamiento por la izquierda y censura por la derecha.
- AFT tiempo de vida acelerada.
- PH riesgos proporcionales.
- $\mathbf{I}_k$ matriz identidad de dimensión k.
- *c.s.* "casi seguro".
- $A \wedge B$ mínimo de A y B.
- $A \vee B$ máximo de A y B.
- $h^* = \underset{h \in H}{\operatorname{argmin}}\{F(h)\}$ indica $F(h^*) = \min_{h \in H}\{F(h)\}$.
- $X_n \xrightarrow[n\to\infty]{c.s.} X$ indica que X_n converge "casi seguro" a X cuando n tiende a infinito.
- $X_n \xrightarrow[n\to\infty]{\mathcal{P}} X$ indica que X_n converge en probabilidad a X cuando n tiende a infinito.
- $X_n \xrightarrow[n\to\infty]{\mathcal{D}} X$ indica que X_n converge en distribución (o en ley) a X cuando n tiende a infinito.

- Para a_n y b_n series de números reales, $a_n = o(b_n)$ indica que $\lim_{b_n \to 0} |a_n/b_n| = 0$.
- Para A_n y B_n series de variables aleatorias reales, $A_n = o_P(B_n)$ indica que $\forall \delta > 0$, $\lim_{n \to \infty} Pr\{|A_n/B_n| > \delta\} = 0$.

Capítulo 1

Conceptos básicos

1.1. Introducción

En este capítulo inicial se presentan algunos conceptos teóricos y herramientas básicas que serán utilizados en esta memoria.

La estructura de este capítulo es la siguiente: En la Sección 1.2 se presentan conceptos relativos a los procesos de recuento, como pueden ser: filtración, tiempo de parada, proceso predecible, proceso de salto, proceso de intensidad y modelo de intensidad multiplicativo, entre otros. La definición de proceso de recuento para el análisis teórico en Análisis de Supervivencia resulta conveniente, ya que permite enlazar con la teoría de martingalas a través de la descomposición de Doob-Meyer. Las propiedades asintóticas de los estimadores se derivan del teorema central del límite para martingalas locales de cuadrado integrable dado en Rebolledo (1980).

En la Sección 1.3 se presenta un enfoque basado en estos procesos de recuento para la modelización de fenómenos que ocurren en el tiempo, tales como los modelos de tiempo de vida (un único evento) o modelos para eventos recurrentes. La formulación se presenta bajo el modelo de intensidad multiplicativo de Aalen (1978). En la Sección 1.4 se define el proceso de Poisson no homogéneo y las propiedades de dicho proceso.

En la Sección 1.5 se estudian dos enfoques diferentes para la estimación de la

función de intensidad de un proceso de recuento. En primer lugar, considerando un enfoque paramétrico, los parámetros se estiman mediante máxima verosimilitud. En segundo lugar, se considera un enfoque no paramétrico y se describe el estimador de la función de azar acumulado de Nelson-Aalen (Nelson (1969, 1972) y Aalen (1975, 1978)).

En la Sección 1.6 se presentan métodos de estimación tipo núcleo para la función de azar basados en procesos de recuento. El método de estimación núcleo se describe en términos generales mediante el estimador de la función de densidad propuesta originariamente por Rosenblatt (1956) y Parzen (1962). Seguidamente, se muestran los estimadores tipo núcleo de la función de intensidad de procesos de recuento de Ramlau-Hansen (1983), y de la función de supervivencia de Kulasekera, Williams, Coffin y Manatunga (2001), así como las propiedades asintóticas de ambos estimadores. A continuación, se presentan los estimadores locales mínimo-cuadráticos de la función de azar, propuestos por Nielsen y Tanggaard (2001), y de la función de densidad, propuestos por Nielsen, Tanggaard y Jones (2009). También se muestran las propiedades asintóticas de los mismos. Para finalizar, se presentan dos métodos clásicos de elección del parámetro de suavizado para el estimador de la intensidad.

1.2. Herramientas probabilísticas

A continuación se recogen algunos conceptos básicos relacionados con la teoría de martingalas que serán necesarios para el desarrollo de este trabajo. Para un estudio más detallado pueden verse por ejemplo los textos de Fleming y Harrington (1991) y Andersen, Borgan, Gill y Keiding (1993).

1.2.1. Conceptos básicos

Sea el espacio de probabilidad $(\Omega, \mathcal{F}, P)$ y supóngase que se fija un intervalo de tiempo continuo

$$\mathcal{T} = [0, \tau) \quad \text{o} \quad [0, \tau]$$

para un tiempo final τ, con $0 < \tau \leq \infty$. Nótese que el tiempo τ puede estar o no incluido en el intervalo, esto variará según la aplicación deseada.

Para fijar conceptos y notación, se muestran las siguientes definiciones.

Definición 1. *Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad. Una **filtración** no decreciente y continua a la derecha $\{\mathcal{F}_t : t \in \mathcal{T}\}$, es una familia continua, creciente a la derecha de sub-σ-álgebras de $\mathcal{F}$.*

Se dice que $\{\mathcal{F}_t\}$ satisface las **condiciones usuales** *(les conditions habituelles)* de Andersen *et al.* (1993), si se verifica

- $\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}, \quad \forall s < t$ (no decreciente)
- $\mathcal{F}_s = \bigcap_{t>s} \mathcal{F}_t, \quad \forall s$ (continua a la derecha)
- $A \subset B \in \mathcal{F}, Pr\{B\} = 0 \Rightarrow A \in \mathcal{F}_0$ (completa)

$\mathcal{F}_t$ es la σ-álgebra que contiene todos los eventos cuya ocurrencia, o no, está determinada en el instante t. $\mathcal{F}_{t-}$, es la menor de las σ-álgebras que contiene a todos las $\mathcal{F}_s$, con $s < t$, es decir, que contiene los eventos determinados estrictamente antes de t.

Definición 2. *Una función Z de Ω en $\mathbb{R}$ se dice que es una variable aleatoria (v.a.) o bien se dice **medible** (relativa a $\mathcal{F}$) si el suceso $\{Z \leq x\} = \{\omega \in \Omega : Z(\omega) \leq x\} \in \mathcal{F}, \forall x$. Esto es, Z es una aplicación medible de $(\Omega, \mathcal{F}, P)$ en $(\mathbb{R}, \mathcal{B})$ donde $\mathcal{B}$ es la σ-álgebra de Borel.*

Definición 3. *Un **proceso estocástico** X es una colección de variables aleatorias $X = \{X(t) : t \in \mathcal{T}\}$ definidas sobre el espacio de probabilidad $(\Omega, \mathcal{F}, P)$.*

Por ejemplo, $X(t)$ puede representar el número de personas que se recuperan de cierta enfermedad hasta el instante t, o la posición de una partícula en el instante t, o el estado de un componente mecánico en el instante t.

Definición 4. *Un proceso estocástico X se dice que es* ***adaptado*** *a la filtración $\mathcal{F}_t$ si $X(t)$ es $\mathcal{F}_t$-medible para cada t. El valor de la v.a. $X(t)$ en el punto $\omega \in \Omega$ se representa como $X(t,\omega)$. Así, el proceso estocástico X no sólo puede verse como una función de ω para un valor fijo de t (una v.a.), sino también puede verse como una función de t para un valor fijo de ω. Esta función $X(\cdot,\omega)$ se llama* ***trayectoria*** *de X.*

Las siguientes tres definiciones son propiedades importantes de integrabilidad y de acotación para los procesos estocásticos. Un proceso estocástico X es

1. integrable si
$$\sup_{0\leq t<\infty} \{E\left[|X(t)|\right]\} < \infty.$$
2. de cuadrado integrable si
$$\sup_{0\leq t<\infty} \left\{E\left[X^2(t)\right]\right\} < \infty.$$
3. acotado si existe una constante finita ξ tal que
$$Pr\left\{\sup_{0\leq t<\infty} \{X(t)\} < \xi\right\} = 1.$$

Definición 5. *El proceso estocástico X se llama* ***càdlàg*** *(de su abreviatura en francés: continu à droite, limité à gauche) si sus trayectorias $\{X(t,\omega) : t \in \mathcal{T}\}$ para casi todos los ω, son continuas a la derecha con límites a la izquierda. Al conjunto de funciones càdlàg se le denota como $D(\mathcal{T})$, dominio que se conoce como espacio de Skorokhod.*

A menudo se describe una filtración como la filtración generada por un proceso estocástico X. Esto significa que $\mathcal{F}_t$ es la σ-álgebra generada por $X(s)$, para $s \leq t$.

También se tiene que $\mathcal{F}_{t-}$ es generada por $X(s)$, para $s < t$. Frecuentemente, se le añaden eventos a $\mathcal{F}_t$ para fijar el instante inicial a cero.

Definición 6. *Un* ***proceso de salto*** X *es un proceso en el que para cada* $t \in [0, \tau]$ *y* $\omega \in \Omega$, $X(s, \omega)$ *es constante en* $s \in [t, t + \varepsilon)$ *para algún* $\varepsilon > 0$ *(que depende de* t *y* ω*, en general).*

Un resultado importante en Courrège y Priouret (1965), establece que la filtración generada por un proceso de salto continuo a la derecha es continua a la derecha.

Definición 7. *Un* ***tiempo de parada*** T *es una v.a. que toma valores en* $[0, \tau]$*, tales que*

$$\{T \leq t\} \in \mathcal{F}_t, \ \forall t \in \mathcal{T}.$$

Un ejemplo importante de tiempo de parada es la primera vez que un suceso supera un valor dado, así si X es *càdlàg* y adaptado entonces $T = \inf\{t : |X(t)| \geq c\}$ es un tiempo de parada.

Definición 8. *Un conjunto de variables aleatorias* $X(T)$*, es* ***uniformemente integrable*** *para todos los tiempos de parada* T *si*

$$\lim_{c \to \infty} \sup_{T} \{E\left[|X(T)| I\left[|X(T)| > c\right]\right]\} = 0.$$

La predictibilidad de un proceso estocástico es un concepto que va a definirse a continuación, aunque para ello antes se necesita conocer la definición de σ-álgebra predecible para una filtración. La idea intuitiva de predictibilidad es simple, un proceso estocástico X es predecible si su comportamiento en el instante t viene determinado por su comportamiento anterior en $[0, t)$.

Definición 9. *Sea* $(\Omega, \mathcal{F}, P)$ *un espacio de probabilidad equipado con una filtración* $\{\mathcal{F}_t : t \in \mathcal{T}\}$*. La* σ*-álgebra sobre* $\mathcal{T} \times \Omega$ *generada por todos los conjuntos de la forma*

$$[0] \times A, \qquad A \in \mathcal{F}_0,$$

y

$$(a, b] \times A, \qquad a < b, \qquad a, b \in \mathcal{T}, \qquad A \in \mathcal{F}_a,$$

se llama ***σ-álgebra predecible*** *para la filtración* $\{\mathcal{F}_t : t \in \mathcal{T}\}$.

Definición 10. *Un proceso estocástico X se dice que es un* ***proceso predecible*** *con respecto a la filtración $\{\mathcal{F}_t : t \in \mathcal{T}\}$, y se suele escribir también que X es $\mathcal{F}_t$-predecible, si es medible con respecto a la σ-álgebra predecible generada por la filtración.*

Definición 11. *Si X es un proceso predecible con respecto a la filtración $\{\mathcal{F}_t : t \in \mathcal{T}\}$, entonces para cualquier $t > 0$, $X(t)$ es $\mathcal{F}_{t-}$-medible.*

Existen varias caracterizaciones equivalentes, por ejemplo, se dice que un proceso estocástico X es predecible si y sólo si $X(T)$ es $\mathcal{F}_{T-}$-medible para todo tiempo de parada T. De este modo, el valor de un proceso predecible en el tiempo T viene determinado por lo que ha ocurrido justo antes de ese tiempo.

La expresión de la integral de un proceso estocástico con respecto a otro, juega un papel importante en el resto de este capítulo.

Definición 12. *Sean dos procesos estocásticos X e Y, la* ***integral estocástica*** *de X con respecto a Y, se representa por $\int X dY$, es la integral de Lebesgue-Stieltjes sobre el intervalo $[0, t]$ para cada $t \in \mathcal{T}$, que viene dada por*

$$t \mapsto \int_0^t X(s)dY(s) = \int_{[0,t]} X(s)dY(s),$$

definida para cada ω y t tal que

$$\int_{[0,t]} |X(s)||dY(s)| < \infty.$$

En este caso, Y se supone que es un proceso *càdlàg* y se dice que es un *proceso de variación finita* cuando $\int_{[0,t]} |dY(s)|$ es finita para todo $t \in \mathcal{T}$, para casi todo $\omega \in \Omega$, y al proceso $\int |dY|$ se le llama proceso de variación.

Definición 13. *Un **proceso de recuento** $\{N(t) : t \in \mathcal{T}\}$ es un proceso estocástico adaptado a la filtración $\{\mathcal{F}_t : t \in \mathcal{T}\}$, de tipo càdlàg con $N(0) = 0$ y $N(t) < \infty$ c.s. y cuyas trayectorias son constantes a trozos con saltos de tamaño unitario.*

Definición 14. *Sea $\{N(t) : t \in \mathcal{T}\}$ un proceso estocástico adaptado a la filtración $\{\mathcal{F}_t : t \in \mathcal{T}\}$, se define la **función de intensidad condicionada** como*

$$\lambda(t) = \lim_{\Delta \to 0} \frac{Pr\{N(t+\Delta) - N(t) \geq 1 | \mathcal{F}_{t^-}\}}{\Delta}. \tag{1.1}$$

siendo Δ una cantidad infinitesimal.

Definición 15. *Sea el proceso de recuento $\{N(t) : t \in \mathcal{T}\}$ adaptado a la filtración $\{\mathcal{F}_t : t \in \mathcal{T}\}$. Se dice que dicho proceso de recuento sigue un **modelo de intensidad multiplicativo** si la intensidad λ existe y puede escribirse de la forma*

$$\lambda(s) = \alpha(s)Y(s), \tag{1.2}$$

siendo α la función de azar subyacente, continua y determinística, e Y el proceso de riesgo predecible con respecto a la filtración.

1.2.2. Algunas nociones sobre teoría de martingalas

Sea $M = \{M(t) : t \in \mathcal{T}\}$ un proceso estocástico de tipo *càdlàg* y $\{\mathcal{F}_t : t \in \mathcal{T}\}$ una filtración, definidos sobre el mismo espacio de probabilidad. Se dice que M es una *martingala* con respecto a la filtración $\mathcal{F}_t$ si se cumple que

1. M es adaptado a $\mathcal{F}_t$,
2. $E\left[|M(t)|\right] < \infty, \ \forall t \in \mathcal{T}$,
3. $E\left[M(t)|\mathcal{F}_s\right] = M(s), \ \forall s \leq t$.

El proceso será una *submartingala* si la tercera condición se sustituye por

$$E\left[M(t)|\mathcal{F}_s\right] \geq M(s), \ \forall s \leq t$$

y será una *supermartingala* si dicha condición es

$$E\left[M(t)|\mathcal{F}_s\right] \leq M(s),\ \forall s \leq t.$$

Una martingala M se dice que es de *cuadrado integrable* si se verifica que

$$\sup_{t\in\mathcal{T}} \left\{E\left[M^2(t)\right]\right\} < \infty.$$

Descomposición de Doob-Meyer

La siguiente descomposición de Doob-Meyer no es una versión general ya que sólo se establece para el caso de submartingalas, no negativas y continuas a la derecha.

Teorema 1. *Sea X una submartingala no negativa y continua a la derecha con respecto a un espacio de probabilidad $(\Omega, \mathcal{F}, P)$ dotado con una filtración continua a la derecha $\{\mathcal{F}_t : t \geq 0\}$. Entonces existe una martingala continua a la derecha M y un proceso predecible continuo a la derecha y no decreciente Λ tal que $E\left[\Lambda(t)\right] < \infty$ y*

$$X(t) = M(t) + \Lambda(t) \qquad c.s. \quad \forall t \in \mathcal{T}.$$

Si $\Lambda(0) = 0$ c.s. y si $X = M' + \Lambda'$ es otra descomposición con $\Lambda'(0) = 0$, entonces para cualquier t se tiene

$$Pr\{M'(t) \neq M(t)\} = 0 = Pr\{\Lambda'(t) \neq \Lambda(t)\}.$$

Además, si X está acotada entonces M es uniformemente integrable y Λ es integrable.

El proceso Λ en la descomposición de Doob-Meyer se llama *compensador* para la submartingala X.

Puesto que cualquier proceso adaptado, no decreciente y no negativo con esperanza finita es una submartingala, se puede formular el siguiente corolario.

Corolario 1. *Sea* $\{N(t) : t \in \mathcal{T}\}$ *un proceso de recuento adaptado a una filtración continua a la derecha* $\{\mathcal{F}_t : t \in \mathcal{T}\}$*, con* $E[N(t)] < \infty$ *para cualquier* t*. Entonces existe un único proceso* $\mathcal{F}_t$*-predecible, continuo a la derecha y no decreciente,* $\Lambda(t)$*, tal que* $\Lambda(0) = 0$ *c.s.,* $E[\Lambda(t)] < \infty$ *para cualquier* t*, y* $\{M(t) = N(t) - \Lambda(t) : t \in \mathcal{T}\}$ *es una* $\mathcal{F}_t$*-martingala continua a la derecha.*

Un proceso de recuento N es una submartingala local y por tanto tiene un compensador Λ. El proceso Λ es no decreciente y predecible con valor cero en el instante inicial tal que

$$M = N - \Lambda$$

es una martingala local con respecto a $\mathcal{F}_t$. De hecho, M es una martingala local de cuadrado integrable y cumple que

$$E[N(t)] = E[\Lambda(t)],$$

y además, si $E[\Lambda(t)] < \infty$ entonces M es una martingala.

El compensador anterior tiene la forma

$$\Lambda(t) = \int_0^t \lambda(s)ds,$$

siendo el *proceso de intensidad* $\lambda(t)$ un proceso predecible.

Un ejemplo ampliamente conocido de proceso de recuento es el proceso de Poisson, donde los saltos suceden de manera aleatoria e independientes unos de otros. Este proceso se describe mediante su intensidad λ que se define como la probabilidad de ocurrencia de un evento en un intervalo pequeño de tiempo dividido por la longitud del intervalo. En general, para modelizar procesos de recuento, es necesario extender esta idea de intensidad. Si se considera un intervalo pequeño de tiempo $[t, t + dt)$ y se supone que a lo sumo puede suceder un evento en dicho intervalo, entonces el proceso de intensidad $\lambda(t)$ se calcula como la probabilidad de que suceda un evento en $[t, t + dt)$, condicionada a la filtración $\{\mathcal{F}_t : t \in \mathcal{T}\}$, es decir,

$$\lambda(t)dt \approx Pr\{dN(t) = 1|\mathcal{F}_{t-}\}, \tag{1.3}$$

donde $dN(t)$ indica el número de saltos del proceso en $[t, t+dt)$.

Si el proceso está ordenado, y $dN(t)$ es una variable binaria, se puede escribir de forma alternativa como

$$\lambda(t)dt \approx E\left[dN(t)|\mathcal{F}_{t-}\right], \tag{1.4}$$

o, equivalentemente $E\left[dN(t) - \lambda(t)dt|\mathcal{F}_{t-}\right] = 0$.

Definiendo

$$M(t) = N(t) - \int_0^t \lambda(s)ds, \tag{1.5}$$

se verifica que

$$E\left[dM(t)|\mathcal{F}_{t-}\right] = 0. \tag{1.6}$$

La expresión (1.6) resulta interesante ya que representa una definición intuitiva de martingala. Por lo tanto, la definición (1.3) de un proceso intensidad es equivalente a afirmar que la diferencia entre el proceso de recuento $N(t)$ y el proceso de intensidad acumulada, $\Lambda(t) = \int_0^t \lambda(s)ds$, es una martingala.

A partir de la expresión (1.5), el incremento $dN(t)$ del proceso de recuento puede escribirse como

$$dN(t) = \lambda(t)dt + dM(t). \tag{1.7}$$

Esta es una relación de la forma "observación=señal+ruido", siendo $dN(t)$ la observación, $\lambda(t)dt$ la señal y $dM(t)$ el término de ruido. De la expresión (1.7) se obtiene que la martingala $M(t) = \int_0^t dM(s)$ es un ruido acumulado. Por tanto, la martingala (1.5) tiene una expresión similar a la suma de errores aleatorios en la estadística clásica.

Teorema Central del Límite para martingalas

Uno de los puntos fuertes de la representación de datos de supervivencia mediante procesos de recuento es la utilización de martingalas y que se dispone de un Teorema Central del Límite para martingalas debido a Rebolledo (1980). Dicho teorema facilita la obtención de las propiedades asintóticas de los estimadores.

Sea $\{N^{(n)}(t) : t \in \mathcal{T}\}$ una secuencia de procesos de recuento con n componentes para cada $n = 1, 2, \ldots$, adaptados a las filtraciones $\{\mathcal{F}_t^{(n)} : t \in \mathcal{T}\}$, también se supone que dichos procesos de recuento satisfacen el modelo de intensidad multiplicativo de Aalen (1.2) con procesos de riesgo $Y^{(n)}(t)$. Sea $Z^{(n)}(t)$ una matriz de procesos estocásticos predecibles localmente acotados con respecto a la filtración anterior $\mathcal{F}_t^{(n)}$, y sea $M^{(n)}(t) = N^{(n)}(t) - \Lambda^{(n)}(t)$ donde $\Lambda^{(n)}(t)$ es el compensador de $N^{(n)}(t)$. Por tanto, el proceso de covarianza predecible de la martingala $\widetilde{M}^{(n)}(t) = \int_0^t Z^{(n)}(s) dM^{(n)}(s)$ es

$$\begin{aligned}\langle \widetilde{M}^{(n)} \rangle(t) &= \int_0^t (Z^{(n)}(s))^2 d\langle M \rangle^{(n)}(s) = \int_0^t (Z^{(n)}(s))^2 d\Lambda^{(n)}(s) \\ &= \int_0^t (Z^{(n)}(s))^2 \alpha(s) Y^{(n)}(s) ds.\end{aligned}$$

Ramlau-Hansen (1983) demostró el siguiente resultado que se utilizará más adelante:

Teorema 2. *Se supone que si*

1. *(Estabilidad asintótica). Si* $\exists \upsilon > 0$ *se tiene*

$$\int (Z^{(n)}(s))^2 \alpha(s) Y^{(n)}(s) ds \xrightarrow[n\to\infty]{\mathcal{P}} \upsilon.$$

2. *(Condición de Lindeberg).* $\forall \varepsilon > 0$ *se tiene*

$$\int (Z^{(n)}(s))^2 I\left[|Z^{(n)}(s)| > \varepsilon\right] \alpha(s) Y^{(n)}(s) ds \xrightarrow[n\to\infty]{\mathcal{P}} 0.$$

Entonces

$$\widetilde{M}^{(n)} \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \upsilon).$$

1.3. El modelo multiplicativo de Aalen

Sea $N(t)$ un procesos de recuento que registra las ocurrencias de un evento de interés en el intervalo $(0, t]$. Este proceso tiene función de intensidad $\lambda(t)$ como se

definió en (1.1). Bajo el modelo de intensidad multiplicativo de Aalen (1978), se asume que la función de intensidad $\lambda(s) = \alpha(s)Y(s)$ se puede factorizar como una parte determinística α, y un factor aleatorio Y, que es un proceso predecible.

Existen numerosas situaciones que pueden ser descritas mediante un proceso de recuento $N(t)$, con proceso de intensidad $\lambda(t)$ de forma multiplicativa, siendo $\alpha(t)$ una función determinística no negativa e $Y(t)$ un proceso estocástico, observable y continuo a la izquierda.

1.3.1. Procesos de recuento con un único salto

Sea T una v.a. que representa un tiempo de supervivencia con función de azar $\alpha(t)$, a esta v.a. no negativa, se le puede asociar el proceso $N(t) = I[T \leq t]$, que toma el valor 1 si el individuo experimenta el evento de interés y 0 en otro caso. El correspondiente proceso de intensidad, se define a partir de

$$Pr\{dN(t) = 1|\mathcal{F}_t\} = Pr\{t \leq T < t + dt|\mathcal{F}_t\} = \begin{cases} \alpha(t)dt, & T \geq t, \\ 0, & T < t, \end{cases}$$

ya que la filtración $\mathcal{F}_t$ contiene toda la información del suceso, es decir, si $T \geq t$ o si $T < t$. Entonces, el proceso de intensidad es $\lambda(t) = \alpha(t)I[T \geq t]$.

Si se consideran $T_1, T_2, \ldots, T_n$ variables aleatorias no negativas e independientes, correspondientes a tiempos de supervivencia de n individuos y $\alpha_i(t)$ la función de azar de T_i, a partir de los tiempos de supervivencia, pueden definirse los procesos de recuento $N_i(t) = I[T_i \leq t]$, $i = 1, 2, \ldots, n$, que tienen procesos de intensidad asociados

$$\lambda_i(t) = \alpha_i(t)I[T_i \geq t], \quad i = 1, 2, \ldots, n. \tag{1.8}$$

A partir de los procesos de recuento *individuales* $N_1(t), N_2(t), \ldots, N_n(t)$ se puede obtener un proceso *agregado* $N(t)$ como la suma de los procesos individuales

$$N(t) = \sum_{i=1}^{n} N_i(t).$$

Este proceso cuenta el número de individuos que han experimentado el evento en el instante t. Para el caso continuo considerado aquí, no puede haber dos T_i iguales. De este modo, el proceso agregado tiene saltos de magnitud 1 en cada T_i, y por lo tanto es un proceso de recuento. Entonces, mediante linealidad del operador esperanza se tiene que

$$E\left[dN(t)|\mathcal{F}_t\right] = \sum_{i=1}^{n} E\left[dN_i(t)|\mathcal{F}_t\right],$$

y de forma análoga a la expresión (1.4), el proceso de intensidad $\lambda(t)$ viene dado por

$$\lambda(t) = \sum_{i=1}^{n} \lambda_i(t). \tag{1.9}$$

Para el caso particular en el que $\alpha_i(t) = \alpha(t)$, $\forall i$, es decir, cuando los tiempos de supervivencia son i.i.d., si se sustituye la expresión (1.8) en la ecuación (1.9), el proceso de recuento agregado tiene un proceso de intensidad multiplicativa dado por la expresión (1.2), siendo $Y(t) = \sum_{i=1}^{n} I\left[T_i \geq t\right] = n - N(t-)$ el número de individuos en riesgo "justo antes" del instante t, donde $N(t-)$ es el valor del proceso de recuento agregado "justo antes" del instante t.

Información muestral incompleta

En la observación de tiempos de vida, dos de los tipos más comunes de información muestral incompleta vienen dados por la presencia de censura por la derecha y truncamiento por la izquierda, entendiendo por la primera, que algunos tiempos exceden una ventana de observación cuyos límites son los tiempos de censura, y entendiendo por la segunda, que un individuo comienza a observarse posteriormente a la ocurrencia del evento inicial. El análisis estadístico estándar para datos completos resulta inadecuado cuando las observaciones están censuradas y/o truncadas, ya que las estimaciones resultantes pueden ser sesgadas.

Una representación gráfica de datos de vida censurados por la derecha y truncados por la izquierda puede verse en la Figura 1.1, donde los sujetos 3 y 8 representan

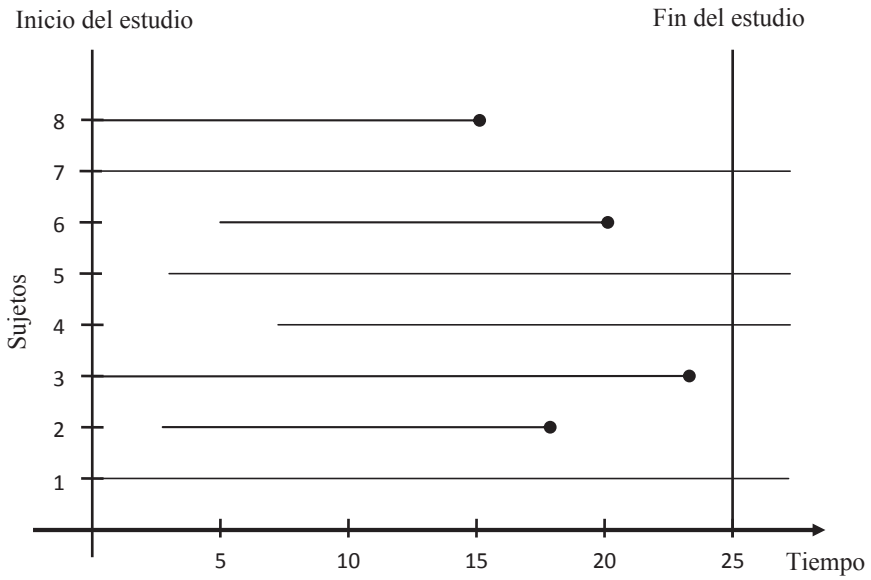


Figura 1.1: Representación gráfica de datos de tiempos de vida: completos, censurados por la derecha, truncados por la izquierda y combinación de datos censurados por la derecha y truncados por la izquierda.

a datos completos, los sujetos 1 y 7 representan a datos censurados por la derecha, los sujetos 2 y 6 representan a datos truncados por la izquierda y los sujetos 4 y 5 representan a datos censurados por la derecha y truncados por la izquierda.

Las definiciones y teoremas presentados en esta sección, han sido tomadas de Martinussen y Scheike (2006).

Se considera un proceso de recuento multivariante definido de la forma $N^*(t) = (N_1^*(t), N_2^*(t), \ldots, N_n^*(t))$ adaptado a la filtración $\mathcal{F}_t^*$ y definido en un espacio de probabilidad $(\Omega, \mathcal{F}, P)$. Se supone que el proceso de recuento $N_i^*(t)$ tal que la información acumulada a lo largo del tiempo es

$$\mathcal{F}_t^* = \vee_{i=1}^n \mathcal{F}_t^{i*},$$

que se compone de elementos de información independientes dados por $\mathcal{F}_t^{i*}$, para $i = 1, 2, \ldots, n$.

Se considera que N_i^* tiene un compensador definido por Λ_i^*, donde

$$\Lambda_i^*(t) = \int_0^t \lambda_i^*(s)ds, \quad i = 1, 2, \ldots, n. \tag{1.10}$$

Como se mencionó antes, N^* habitualmente no se observa por completo, sino que sólo se dispone de una versión incompleta. La parte observable de $N_i^*(t)$ puede expresarse de la forma

$$N_i(t) = \int_0^t C_i(s)dN_i^*(s),$$

donde $C_i(t) = I\left[t \in A_i\right]$.

Para simplificar, se supone que el proceso de filtración es independiente a través de los sujetos. El ejemplo principal de filtración es la censura por la derecha, donde $A_i = [0, U_i]$ siendo U_i un tiempo aleatorio, esto es,

$$C_i(t) = I\left[t \leq U_i\right]. \tag{1.11}$$

En este caso, $N_i^*(t)$ es observado sólo hasta el tiempo de censura U_i, para $i = 1, 2, \ldots, n$ y a partir de entonces es desconocido.

La filtración $\mathcal{F}_t^*$ contiene la información con la que se quiere construir el modelo. Desafortunadamente no se puede observar $\mathcal{F}_t^*$ por completo debido a la existencia de información muestral incompleta representada por C.

Se supone que la probabilidad de un salto para el proceso no observado dada la información completa y el proceso observado son equivalentes, esto es, aquellos sujetos en riesgo y en observación son representativos de toda la muestra.

A continuación se consideran dos casos especiales de información muestral incompleta, la censura por la derecha y el truncamiento por la izquierda.

Caso 1. Datos de supervivencia con censura por la derecha

Sea T^* el tiempo de supervivencia de interés y sea $N^*(t) = I\left[T^* \leq t\right]$. Se supone que la distribución de T^* es absolutamente continua con función de azar $\alpha(t)$. Bajo el modelo de intensidad multiplicativo de Aalen se verifica que

$$\lambda^*(t) = \alpha(t)Y^*(t)$$

con respecto a $\mathcal{F}_t^* = \mathcal{F}_t^{N^*}$, donde $Y^*(t) = I\,[t \leq T^*]$ es el indicador de riesgo. Sea U el tiempo de censura, que se supone independiente del tiempo de fallo. Sólo se observa el tiempo de fallo T^* si no excede al correspondiente tiempo de censura U, y por tanto se observa

$$T = T^* \wedge U \quad \text{y} \quad \delta = I\,[T \leq U]\,.$$

La condición de censura independiente puede escribirse como

$$Pr\{t \leq T^* < t + dt | T^* \geq t\} = Pr\{t \leq T^* < t + dt | T^* \geq t, U \geq t\}. \tag{1.12}$$

Normalmente, la condición (1.12) se utiliza como la definición de censura por la derecha independiente, (ver por ejemplo Fleming y Harrington (1991) p. 27). La filtración observada bajo la censura considerada viene dada por

$$\mathcal{F}_t = \sigma\left\{(N(s), Y(s+)) : 0 \leq s \leq t\right\},$$

donde $Y(t) = C(t)Y^*(t)$ y $C(t) = I\,[t \leq U]$. Ya que la censura es independiente, el proceso de recuento observado $N(t)$ tiene proceso de intensidad

$$\lambda(t) = \alpha(t)Y(t) \tag{1.13}$$

con respecto a $\mathcal{F}_t$. En la expresión (1.13) puede verse que la estructura de intensidad multiplicativa queda preservada y que la parte determinística queda inalterada. La única diferencia es que el indicador de riesgo $Y^*(t) = I\,[t \leq T]$ se reemplaza por el indicador de riesgo observado $Y(t) = C(t)Y^*(t) = I\,[t \leq T \wedge C]$. Esto es importante porque por ejemplo, para realizar estimaciones por máxima verosimilitud en datos de supervivencia con censura por la derecha, se supone que la censura por la derecha no incluye ninguna información sobre el parámetro de interés.

□

Caso 2. Datos de supervivencia con truncamiento por la izquierda

Sea el proceso de recuento multivariante $N^*(t) = (N_1^*(t), N_2^*(t), \ldots, N_n^*(t))$ adaptado a la filtración $\mathcal{F}_t^*$ y que en el conjunto de datos de tiempos de fallo

puede definirse $N^*(t)$ a partir de tiempos de fallo i.i.d. $T_1^*, T_2^*, \ldots, T_n^*$ tal que $N_i^*(t) = I\left[T_i^* \leq t\right]$. Se supone que N^* tiene compensador Λ^* donde

$$\Lambda_i^*(t) = \int_0^t \lambda_i^*(s)ds.$$

Además de los tiempos observados también se dispone de los tiempos de truncamiento $L_1, L_2, \ldots, L_n$ i.i.d. tales que $T_i^* > L_i$.

La diferencia más importante con respecto al caso de censura por la derecha es que los eventos se ven de manera condicionada en el evento de truncamiento.

El proceso de recuento observado puede escribirse de la forma

$$N_i(t) = \int_0^t C_i(s)dN_i^*(s),$$

donde $C_i(t) = I\left[t \geq L_i\right]$.

La filtración independiente es a menudo más complicada en estos conjuntos de datos debido a que la medida de probabilidad relevante está condicionada al evento $A = \cap_{i=1,2,\ldots,n}(T_i^* > L_i)$. En este caso también se supone que el truncamiento es independiente del proceso de fallo y entonces

$$\begin{aligned} C_i(t)Pr\{t \leq T_i^* < t + dt | T_i^* \geq t\} \\ = Pr\{t \leq T_i^* < t + dt | T_i^* \geq t,\, T_i^* \geq L_i,\, L_i\}. \end{aligned} \tag{1.14}$$

□

La censura por la derecha y el truncamiento por la izquierda a menudo se combinan en estudios de supervivencia. Esto corresponde a un proceso de la forma $C_i(t) = I\left[L_i \leq t \leq U_i\right]$ en el que la intensidad se observa sujeta a la información contenida en $\mathcal{F}_{L_i}^{i*}$.

Una representación gráfica de los procesos $N(t)$ y $Y(t)$ puede verse en la Figura 1.2, en el gráfico de la izquierda se representan datos completos, en el gráfico central se representan datos censurados por la derecha y en el gráfico de la derecha se representan datos truncados por la izquierda.

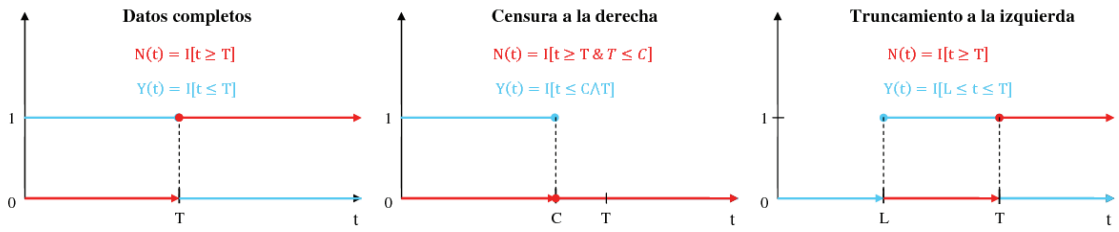


Figura 1.2: Representación gráfica de los procesos $N(t)$ e $Y(t)$, para datos completos, censura por la derecha y truncamiento por la izquierda.

1.3.2. Procesos de recuento para la modelización de eventos recurrentes

En secciones anteriores se ha considerado que el evento de interés ocurre únicamente una vez a lo largo de la vida del sujeto (muerte de un individuo, fallo no reparable, etc.), en esta sección se considera el caso en el que el suceso de interés puede presentarse repetidamente en el tiempo. Este tipo de datos surgen en muchos campos, por ejemplo en el análisis de Fiabilidad, el historial de las reparaciones de un objeto manufacturado puede ser considerado como un suceso recurrente. En Bioestadística los tiempos de recaída de episodios de una enfermedad en pacientes crónicos también pueden ser considerados como eventos recurrentes. En resumen, los datos de eventos recurrentes pueden representar a los tiempos de sucesivas repeticiones de un evento soportados por cada unidad muestral (pacientes de una enfermedad, sistemas en Fiabilidad, etc.). En este tipo de fenómenos el interés se centra en el análisis de la razón de ocurrencia de los sucesos a lo largo del tiempo.

Sea un proceso puntual $N(t)$ observado en un intervalo $[0, t]$ donde $N(t)$ representa el número de eventos que ocurren en dicho intervalo, en general se supone que hay n individuos bajo observación de modo que $N_i(t)$, $i = 1, 2, \ldots, n$, cuenta el numero de ocurrencia del evento de interés para el individuo i-ésimo en el intervalo $[0, t]$. La función de intensidad del proceso N_i se define como la razón de ocurrencia

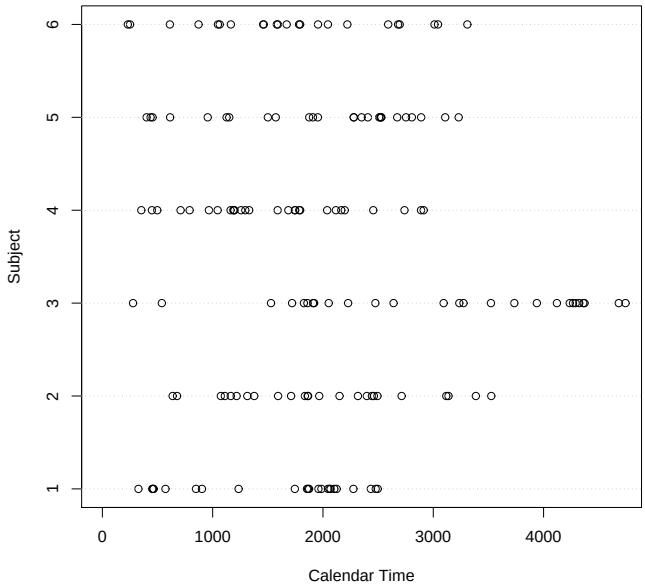


Figura 1.3: Datos de eventos recurrentes: tiempos de fallo (horas) del sistema hidráulico de seis maquinas de LHD.

del suceso dado la historia del proceso hasta t, es decir,

$$\lambda_i(t) = \lim_{\Delta \to 0} \frac{Pr\{N_i(t+\Delta) - N_i(t) > 0 | \mathcal{F}_t^i\}}{\Delta}$$

donde $\mathcal{F}_t^i = \sigma\{N_i(u) : 0 \leq u \leq t\}$ y siendo Δ una cantidad infinitesimal. Una hipótesis esencial es que los sujetos observados son independientes entre sí.

Ejemplo: Un sistema de Fiabilidad

A modo de ilustración se consideran unos datos de Fiabilidad relativos a un sistema hidráulico de máquinas de carga, acarreo y descarga (LHD). Los datos son tomados de Kumar y Klefsjö (1992) y consisten en tiempos entre fallos sucesivos (en horas, excluyendo reparaciones o tiempos de paradas) de sistemas hidráulicos de seis máquinas de LHD, ver Figura 1.3.

□

Siguiendo la formulación de las secciones anteriores, se considera que, bajo el modelo multiplicativo de Aalen,

$$\lambda_i(t) = \alpha_i(t)Y_i(t), \tag{1.15}$$

donde Y_i es el proceso de riesgo que toma el valor 1 si el individuo está en riesgo y en observación, en el instante t y 0 en otro caso, $\alpha_i(t)\Delta$ representa la probabilidad de que el individuo i-ésimo registre el evento de interés en un intervalo de tiempo $[t, t+\Delta)$, y dado que el individuo está en riesgo "justo antes" del instante t.

En general la expresión de $\alpha_i(t)$ puede depender explícitamente del pasado y en este sentido pueden considerarse dos casos extremos. Si $\alpha_i(t)$ es una función determinística que sólo depende de la longitud de tiempo transcurrida desde el instante 0, los datos recurrentes pueden ser modelizados por un proceso de Poisson (más detalles pueden verse en la siguiente sección). Si por otro lado $\alpha_i(t)$ depende del tiempo transcurrido desde el último evento los eventos recurrentes pueden ser modelizados por un proceso de renovación, en este caso la función de intensidad no se ajusta al modelo de intensidad multiplicativo de Aalen y por tanto no será objeto de estudio en este trabajo.

1.4. El proceso de Poisson no homogéneo (PPNH)

Un caso de especial importancia es el PPNH, se incluye aquí como un caso particular de proceso de recuento con función de intensidad que se ajusta al modelo multiplicativo de Aalen.

Antes de comenzar con esta sección, es necesario definir el concepto de *modelos de reparación mínima* que consiste en observar un sistema hasta que sucede un fallo (evento de interés) y cuando se detenga el sistema, se restaura inmediatamente para ponerlo en funcionamiento. Después de la reparación, el estado del sistema es exactamente el mismo que había "justo antes" del fallo.

Se considera un sistema en el que se producen múltiples fallos a lo largo del

tiempo, por lo que se precisa un modelo estocástico para describir la ocurrencia de dichos eventos a través del tiempo (fallos del sistema), es decir, un proceso puntual. Para ello se supone el concepto de reparación mínima, en donde el tiempo transcurrido entre eventos sucesivos no es i.i.d.

A continuación, se establecen algunos conceptos que se utilizarán en lo sucesivo. Se considera que el tiempo varía entre 0 y τ, es decir, $0 = T_0 < T_1 < \ldots < T_n = \tau$ son los tiempos aleatorios ordenados según va sucediendo el evento.

Definición 16 (Proceso de Poisson no homogéneo). *Sea $N(t)$ un proceso de recuento, dicho proceso será un PPNH si satisface las siguientes hipótesis de probabilidad básicas:*

1. *$N(0) = 0$, es decir, el proceso de recuento en el instante inicial es 0.*

2. *$\{N(t), t \geq 0\}$ tiene la propiedad de incrementos independientes, que significa que para cualquier $t_0 < t_1 < t_2 < \ldots < t_n$ entonces*

$$\begin{aligned} N(t_0, t_1) &= N(t_1) - N(t_0), \\ N(t_1, t_2) &= N(t_2) - N(t_1), \\ &\vdots \\ N(t_{n-1}, t_n) &= N(t_n) - N(t_{n-1}), \end{aligned}$$

   son variables aleatorias independientes,

3. *$Pr\{N(t + \Delta) - N(t) = 1\} = \rho(t)\Delta + o(\Delta)$, es decir, conforme el intervalo $[t, t + \Delta]$ se va haciendo más pequeño, la probabilidad de que sólo ocurra un evento es $\rho(t)\Delta$.*

4. *$Pr\{N(t+\Delta) - N(t) \geq 2\} = o(\Delta)$, es decir, conforme el intervalo $[t, t+\Delta]$ se va haciendo más pequeño, la probabilidad de que ocurran dos o más eventos simultáneos es 0.*

En la definición anterior, $o(\Delta)$ denota una cantidad que tiende a cero para valores pequeños de Δ y $\rho(t)$ es la función de intensidad (no condicionada) que se calcula de la forma

$$\rho(t) = \lim_{\Delta \to 0} \frac{Pr\{N(t+\Delta) - N(t) = 1\}}{\Delta}.$$

Esta expresión indica la probabilidad de fallo en un pequeño intervalo de tiempo dividido por la longitud de dicho intervalo. Si $\rho(t)$ es grande, se espera que haya mayor ocurrencia de fallos, mientras que si $\rho(t)$ es pequeño, se espera lo contrario. La función de intensidad es la probabilidad no condicionada de fallo en un pequeño intervalo de tiempo dividido por la longitud del intervalo. Además, en la definición de función de intensidad, no se dice nada sobre la historia del sistema hasta el instante t, el único interés se centra en la ocurrencia del próximo fallo, no siendo necesariamente el primero.

A continuación, se introducen dos funciones clave para describir la evolución de sistemas reparables.

Definición 17 (**Función media de un proceso de recuento).** *Sea* $\{N(t); t \geq 0\}$ *un proceso de recuento, la función media del proceso viene definida por la expresión*

$$m(t) = E[N(t)], \quad t \geq 0.$$

Por tanto, $m(t)$ es una función no decreciente, continua a la derecha que representa el número esperado de fallos hasta el instante t (ver por ejemplo, Rigdon y Basu (2000)).

La siguiente definición proporciona una medida importante para caracterizar algunos tipos de procesos de recuento.

Definición 18 (**Tasa de ocurrencia de fallos (ROCOF)).** *Sea* $m(t)$ *la función media de un proceso de recuento diferenciable, entonces se define la función*

$$\eta(t) = \frac{d}{dt} m(t), \quad t \geq 0.$$

Esta función normalmente es interpretada como la tasa instantánea de cambio en el número de fallos. Cuando la probabilidad de fallos simultáneos es cero, esto es, cuando el proceso es *ordenado* (como sucede en el caso de PPNHs) puede verse que la función de intensidad es igual a la función ROCOF, es decir, $\rho(t) = \eta(t)$.

El PPNH es una clase especial de proceso de recuento. Es importante destacar que, en casos generales, un proceso de recuento no está completamente descrito en términos de la función de intensidad no condicionada (la función ROCOF). Para ello, se necesita una información más completa (función de intensidad condicionada) dada en la Definición 1.1.

Sea $\mathcal{F}_{t^-}$ la historia de un proceso hasta el instante t pero sin incluirlo. $\mathcal{F}_{t^-}$ se genera mediante un conjunto $\{N(s) : 0 \leq s < t\}$ que contiene toda la información sobre los tiempos de fallo en el pasado, es decir, hasta t. La función de intensidad condicionada viene dada por (1.1), esta función, no debe ser confundida con la función ROCOF $\eta(t)$ ni con la función de intensidad no condicionada $\rho(t)$. Sin embargo, en el caso particular de un PPNH, debido a la propiedad de incrementos independientes, la intensidad condicionada depende de la historia del proceso sólo a través del tiempo t, esto es, la función $\lambda(t)$ es una función determinística y se cumple que $\lambda(t) = \eta(t)$.

Bajo la condición de reparación mínima, la función de intensidad condicionada $\lambda(t)$ del proceso PPNH de fallos es igual a la función de intensidad no condicionada $\rho(t)$, que a su vez es igual a la función ROCOF $\eta(t)$ y que también proporciona la función de intensidad de una v.a. que representa el primer tiempo de fallo T_1 del sistema, por lo que, en definitiva se tiene

$$\lambda(t) = \rho(t) = \eta(t) = \alpha(t). \tag{1.16}$$

De ahora en adelante, se utilizará la notación $\lambda(t)$ para esta función.

En un sistema bajo la condición de reparación mínima, la distribución de un PPNH y, consecuentemente, su comportamiento en fiabilidad a lo largo del tiempo, se caracteriza por completo mediante la función ROCOF. Esta función puede tener

diversas formas, por ejemplo, forma de curva de bañera, de bañera invertida, o incluso, con comportamientos cíclicos con múltiples periodicidades o con marcadas tendencias asimétricas. En este sentido, el trabajo de Krivtsov (2007) considera alternativas a los modelos clásicos *log-lineal* y de *ley de potencia*. Krivtsov propone el uso de formas paramétricas correspondientes a un gran número de familias de distribuciones tradicionales para la función de distribución subyacente del proceso. Sin embargo, los modelos paramétricos por lo general, no son lo suficientemente flexibles como para tratar todas las posibles formas de la función de intensidad y por lo tanto los procedimientos no paramétricos son una alternativa más atractiva.

1.5. Inferencia en procesos de recuento

En esta sección se considera, en primer lugar, la inferencia bajo un enfoque paramétrico y, en segundo lugar, un enfoque no paramétrico.

1.5.1. Enfoque paramétrico

Se considera un proceso de recuento multivariante $N = (N_1, N_2, \ldots, N_k)$ con proceso de intensidad $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_k)$ conocido salvo por un parámetro $\beta = (\beta_1, \beta_2, \ldots, \beta_q) \in \mathbf{B}$, donde $\mathbf{B}$ es un conjunto abierto de espacios Euclídeos q-dimensionales. De este modo puede escribirse que $\lambda = \lambda(\beta)$ y por tanto la función de verosimilitud para un proceso de recuento observado hasta el instante $t = \tau$ viene dada por

$$\begin{aligned}\mathcal{L}(\beta) &= \left\{\prod_{t_i \leq \tau} \lambda(t_i)\right\} \exp\left(-\int_0^\tau \lambda(s)ds\right) \\ &= \left\{\prod_{t_i \leq \tau} \exp\left(-\int_{t_{i-1}}^{t_i} \lambda(s)ds\right) \lambda(t_i)\right\} \exp\left(-\int_{t_{N(\tau)}}^{\tau} \lambda(s)ds\right).\end{aligned}$$

Cada término indica la probabilidad de eventos en $[t_{i-1}, t_i)$ y después de experimentar un evento en el tiempo t_i condicionado al pasado del proceso. El último

término especifica la probabilidad de no experimentar eventos desde el último salto hasta el final del período de observación condicionado al pasado del proceso de recuento. La función de verosimilitud se escribe de la forma

$$\begin{aligned}\mathcal{L}(\beta) &\propto \left\{\prod_{t_i\leq\tau} d\Lambda(t_i)\right\}\exp\left(-\int_0^\tau d\Lambda(s)\right)\\ &= \prod_g\prod_{s\leq\tau}(\lambda_g(s))^{\Delta N_g(s)}\exp\left(-\int_0^\tau \lambda^{(g)}(s)ds\right),\end{aligned} \tag{1.17}$$

donde $\Delta N(s) = N(s) - N(s-)$, y $\lambda^{(g)}(s) = \sum_g \lambda_g(s)$. La log-verosimilitud hasta el instante τ para un proceso de recuento multivariante se puede escribir entonces como

$$l(\beta) = \log(\mathcal{L}(\beta)) = \sum_g\left[\int_0^\tau \log(\lambda_g(s))dN_g(s) - \int_0^\tau \lambda_g(s)ds\right],$$

lo que implica que el proceso *score* tiene la forma

$$U(\beta) = \frac{\partial}{\partial\beta}l(\beta) = \sum_g\left[\int_0^\tau \frac{\partial}{\partial\beta}\log(\lambda_g(s))dN_g(s) - \int_0^\tau \frac{\partial}{\partial\beta}\lambda_g(s)ds\right].$$

La función *score* evaluada en el verdadero valor del parámetro β_0, y bajo ciertas condiciones de regularidad, se puede escribir como

$$U(\beta_0) = \sum_g\int_0^\tau \frac{\partial}{\partial\beta}\log(\lambda_g(s))dM_g(s),$$

siendo por tanto una martingala si $\frac{\partial}{\partial\beta}\log(\lambda_g(\tau))$, $g = 1, 2, \ldots, k$, son procesos predecibles y están localmente acotados. En la expresión anterior,

$$M_g(s) = N_g(s) - \int_0^s \lambda_g(u, \beta_0)du.$$

Dadas observaciones de n procesos de recuento $N_i(t)$, $i = 1, 2, \ldots, n$, el estimador de máxima verosimilitud $\widehat{\beta}$ se calcula resolviendo la siguiente ecuación

$$U_n(\beta) = \sum_{i=1}^n\int \frac{\partial}{\partial\beta}\log(\lambda_i(s))dN_i(s) - \int \frac{\partial}{\partial\beta}\lambda_i(s)ds = 0.$$

Bajo las condiciones de regularidad establecidas en el Capítulo VI.1.1. de Andersen *et al.* (1993), $\sqrt{n}(\widehat{\beta} - \beta_0)$ es asintóticamente normal con varianza $\mathcal{I}^{-1}(\beta_0)$, donde los elementos j, k de la matriz de información $\mathcal{I}$ vienen dados como la esperanza de la derivada de segundo orden de la log-verosimilitud con signo negativo, evaluada en el verdadero valor del parámetro β_0. Por tanto, el elemento j, k de la matriz de información es

$$\mathcal{I}_{j,k}(\beta) = E\left[-\frac{\partial^2}{\partial\beta_j\partial\beta_k} l_n(\beta)\right],$$

evaluado en β_0. La matriz de información puede estimarse de manera consistente mediante la matriz de información observada con elementos

$$I_{j,k}(\beta) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \frac{\partial^2}{\partial\beta_j\partial\beta_k}\log(\lambda_i(s))dN_i(s) + \int_0^\tau \frac{\partial^2}{\partial\beta_j\partial\beta_k}\lambda_i(s)ds,$$

evaluados en $\widehat{\beta}$. Para un estudio más detallado ver Borgan (1984).

1.5.2. Enfoque no paramétrico

Sea la historia de un proceso dada por la filtración $\mathcal{F}_t$, y sea $\Lambda(t)$ el compensador del proceso de recuento $N(t)$. El compensador viene dado por la expresión

$$\Lambda(t) = \int_0^t \lambda(s)ds, \tag{1.18}$$

y bajo el modelo multiplicativo de Aalen, $N(t)$ tiene asociado el proceso de intensidad

$$\lambda(t) = \alpha(t)Y(t).$$

El proceso $\Lambda(t)$ es $\mathcal{F}_t$-adaptado y continuo a la izquierda, y por tanto es predecible. Se tiene además que $dN(t)$ es una variable de Bernoulli con probabilidad condicionada $\alpha(t)Y(t)dt$ dada la filtración $\mathcal{F}_{t-}$, por tanto

$$E\left[dN(t)|\mathcal{F}_{t-}\right] = \alpha(t)Y(t)dt = d\Lambda(t) = E\left[d\Lambda(t)|\mathcal{F}_{t-}\right],$$

lo que justifica la expresión (1.6) de las martingalas para $M = N - \Lambda$.

Sea $\{N_i(t), Y_i(t);\ i = 1, 2, \ldots, n\}$, réplicas del modelo anterior y sea $\mathcal{F}_t = \vee_i \mathcal{F}_t^i$, se define

$$N^{(n)}(t) = \sum_{i=1}^{n} N_i(t), \qquad Y^{(n)}(t) = \sum_{i=1}^{n} Y_i(t).$$

El proceso de recuento $N^{(n)}(t)$ tiene compensador

$$\Lambda(t) = \int_0^t \alpha(s) Y^{(n)}(s) ds,$$

y por tanto

$$M^{(n)}(t) = N^{(n)}(t) - \Lambda(t)$$

es una martingala local de cuadrado integrable con respecto a la filtración $\mathcal{F}_t$. En la expresión anterior, $M^{(n)}(t) = \sum_{i=1}^{n} M_i(t)$ con $M_i(t) = N_i(t) - \Lambda_i(t)$, $i = 1, 2, \ldots, n$.

Nuevamente, descomponiendo el proceso de recuento en una martingala y en su compensador se tiene

$$N^{(n)}(t) = \int_0^t \alpha(s) Y^{(n)}(s) ds + M^{(n)}(t)$$

y por tanto $dM^{(n)}(t)$ es un proceso de media cero, esto motiva la ecuación de estimación

$$Y^{(n)}(t) dA(t) = dN^{(n)}(t),$$

donde $A(t) = \int_0^t \alpha(s) ds$. A partir de aquí, Nelson (1969, 1972) y Aalen (1975, 1978) proponen el estimador de Nelson-Aalen definido de la forma

$$\widehat{A}(t) = \int_0^t \frac{J^{(n)}(s)}{Y^{(n)}(s)} dN^{(n)}(s)$$

que es un estimador de la función de azar integrada $A(t)$, donde $J^{(n)}(s) = I\left[Y^{(n)}(s) > 0\right]$, adoptando el criterio $0/0 = 0$.

De este modo, $\widehat{A}(t)$ puede descomponerse como

$$\widehat{A}(t) = \int_0^t J^{(n)}(s) dA(s) + \int_0^t \frac{J^{(n)}(s)}{Y^{(n)}(s)} dM^{(n)}(s).$$

El segundo término de la descomposición anterior es una martingala local de cuadrado integrable, por tanto, $\widehat{A}(t)$ es un estimador insesgado de

$$\int_0^t \alpha(s) Pr\{Y^{(n)}(s) > 0\} ds,$$

lo que indica que (bajo condiciones apropiadas) el estimador de Nelson-Aalen posee buenas propiedades asintóticas. El estimador $\widehat{A}(t)$ para modelos con procesos de recuento dado por Aalen (1975, 1978) generaliza al estimador propuesto por Nelson (1969, 1972).

Propiedades asintóticas del estimador de Nelson-Aalen

Utilizando el teorema central del límite para martingalas pueden derivarse las propiedades asintóticas del estimador de Nelson-Aalen.

Se define

$$\widetilde{A}(t) = \int_0^t J^{(n)}(s) dA(s),$$

de modo que

$$\widehat{A}(t) - \widetilde{A}(t) = \int_0^t \frac{J^{(n)}(s)}{Y^{(n)}(s)} dM^{(n)}(s)$$

es una martingala local de cuadrado integrable, donde $M^{(n)}(t) = \sum_{i=1}^n M_i(t)$ con $M_i(t) = N_i(t) - \int_0^t \alpha(s) Y_i(s) ds$. El error de estimación puede expresarse como suma de un término variable $\widehat{A}(t) - \widetilde{A}(t)$ y otro término estable $\widetilde{A}(t) - A(t)$ descomponiéndose de la forma

$$\begin{aligned} \sqrt{n}(\widehat{A}(t) - A(t)) &= \sqrt{n}\left((\widehat{A}(t) - \widetilde{A}(t)) + (\widetilde{A}(t) - A(t))\right) \\ &= \sqrt{n}\int_0^t \frac{J^{(n)}(s)}{Y^{(n)}(s)} dM^{(n)}(s) + \sqrt{n}\int_0^t (J^{(n)}(s) - 1)\alpha(s) ds. \end{aligned} \tag{1.19}$$

Bajo condiciones de regularidad adecuadas, la distribución asintótica del estimador de Nelson-Aalen en el intervalo $[0, t]$, $t \in \mathcal{T}$ es una martingala Gaussiana, si se cumple que el primer sumando de la expresión (1.19) converge en probabilidad a una martingala Gaussiana, y que el segundo sumando converge a cero en probabilidad.

Siguiendo Andersen *et al.* (1993), p. 190, se supone que $\int_0^t \alpha(s)ds < \infty$, $\forall t \in \mathcal{T}$ y que existe una función positiva $\gamma(s)$ tal que

$$\sup_{s\in[0,t]} \left\{ \left| \frac{Y^{(n)}(s)}{n} - \gamma(s) \right| \right\} \xrightarrow[n\to\infty]{\mathcal{P}} 0.$$

Puede demostrarse entonces que

$$\sup_{s\in[0,t]} \left\{ \left| \sqrt{n} \int_0^s (J^{(n)}(u) - 1)\alpha(u)du \right| \right\} \xrightarrow[n\to\infty]{\mathcal{P}} 0,$$

y ahora centrándose en el término de la martingala $M(s) = \sqrt{n}(\widehat{A}(s) - \widetilde{A}(s))$ se tiene que $\forall s \leq t$,

$$\langle M \rangle(s) = n \int_0^s \frac{J^{(n)}(u)}{Y^{(n)}(u)} \alpha(u)du \xrightarrow[n\to\infty]{\mathcal{P}} v(s)$$

y

$$\langle M_\varepsilon \rangle(s) = n \int_0^s \frac{J^{(n)}(u)}{Y^{(n)}(u)} \alpha(u) I \left[\left| \sqrt{n} \frac{J^{(n)}(u)}{Y^{(n)}(u)} \right| > \varepsilon \right] du \xrightarrow[n\to\infty]{\mathcal{P}} 0.$$

Por tanto,

$$\sqrt{n}(\widehat{A}(s) - A(s)) \xrightarrow[n\to\infty]{\mathcal{D}} G(s)$$

en $D([0,t])$, con $t \in \mathcal{T}$, donde G es una martingala Gaussiana con función de varianza

$$v(s) = \int_0^s \frac{\alpha(u)}{\gamma(u)} du.$$

1.6. Estimación tipo núcleo

En esta sección se consideran técnicas de estimación suave de las características que describen un proceso de recuento.

1.6.1. Estimador tipo núcleo y criterios de error

Esta sección está centrada en la estimación tipo núcleo de la función de densidad.

El estimador no paramétrico más simple para la función de densidad es el histograma. Se trata de un estimador discontinuo de la densidad cuyo comportamiento depende de la elección arbitraria de un punto inicial y de un parámetro ventana, que define la amplitud de los intervalos o clases. Para solventar el problema de la dependencia del punto inicial, Fix y Hodges (1951), introdujeron un estimador naive en un informe no publicado, este estimador sin embargo sigue siendo discontinuo y depende de la ventana. Rosenblatt (1956) y Parzen (1962) propusieron el estimador tipo núcleo, que sí es continuo y que por lo tanto, en la mayor parte de las ocasiones, se ajusta mejor a la realidad de los modelos estudiados, aunque también depende en gran medida de la elección de un parámetro ventana.

El estimador de tipo núcleo de la función de densidad pertenece a un tipo de estimadores denominados estimadores no paramétricos. Estos estimadores se expresan como medias ponderadas de las observaciones que caen dentro de un entorno del punto de estimación.

A partir de una muestra de $X_1, X_2, \ldots, X_n$, v.a.i.i.d. de una variable de interés X con función de densidad f. Rosenblatt (1956) y Parzen (1962) propusieron el siguiente estimador no paramétrico de la función de densidad

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{1.20}$$

siendo $h > 0$ el parámetro de suavizado y K una función núcleo, que habitualmente se supone no negativa y simétrica con respecto a 0, verificando además las siguientes propiedades

$$\int_{-\infty}^{\infty} K(u)du = 1, \quad \int_{-\infty}^{\infty} uK(u)du = 0, \quad \int_{-\infty}^{\infty} u^2K(u)du < \infty$$

En lo sucesivo se denotará por $\mu_j(K) = \int_{-\infty}^{\infty} u^j K(u)du$ a la función de momentos del núcleo y $R(K) = \int_{-\infty}^{\infty} K^2(u)du$ a la función de curvatura del núcleo K.

El estimador (1.20) tiene dos elementos fundamentales. Por una parte está el parámetro de suavizado h, cuyo papel es el de controlar el grado de suavizado

del estimador resultante. Una mala elección del parámetro de suavizado puede derivar en un estimador demasiado ruidoso, tanto infrasuavizado (h pequeño) o sobresuavizado (h grande) ocultando las características de la densidad subyacente. La función núcleo K, habitualmente es fijada por el investigador antes de iniciar el estudio, y define las ponderaciones que se asignan a cada observación en el entorno local considerado. El papel de esta función es menos relevante que el del parámetro de suavizado como se verá más adelante.

El estimador (1.20) puede describirse bajo la formulación de los procesos de recuento antes presentada como sigue. Asociada a cada v.a. X_i, i.i.d. puede definirse un proceso de recuento $N_i(t) = I[t \geq X_i]$ y un proceso de riesgo $Y_i(t) = I[t \leq X_i]$, además puede obtenerse un estimador de la función de distribución asociada a X, de la forma

$$\widehat{F}(t) = \frac{1}{n} \sum_{i=1}^{n} I\left[N_i(t) = 1\right].$$

A partir de aquí, un estimador de la densidad para procesos de recuento puede expresarse como

$$\widehat{f}_h(t) = \frac{1}{h} \int_{-\infty}^{\infty} K\left(\frac{t-s}{h}\right) d\widehat{F}(s)$$

suavizando los saltos del estimador empírico de F.

Para comparar diversos procedimientos de estimación es necesario disponer de mecanismos que midan la bondad de ajuste de los mismos, comúnmente denominados criterios de error. Uno de los más habituales es el error cuadrático integrado (ISE), definido como

$$\mathrm{ISE}(\widehat{f}) = \int \left(\widehat{f}(x) - f(x)\right)^2 dx.$$

La principal característica del ISE es que es un criterio de error global, es decir, que no depende del punto en el que se evalúa el estimador. Sin embargo, este criterio sí depende de la muestra de datos, y hace que se introduzca una cierta variabilidad intrínseca a la propia muestra pero no al estimador. Por esta razón, se define el error cuadrático medio integrado (MISE), que suprime la aleatoriedad procedente

de cada muestra individual mediante el promedio de los resultados obtenidos. Se define

$$\mathrm{MISE}(\widehat{f}) = E\left[\mathrm{ISE}(\widehat{f})\right] = E\left[\int \left(\widehat{f}(x) - f(x)\right)^2 dx\right] = \int E\left[\left(\widehat{f}(x) - f(x)\right)^2\right] dx.$$

Por tanto, el MISE es un criterio de error global que no depende de la muestra empleada.

Por regla general, no es posible obtener una expresión exacta del MISE para el estimador (1.20). Mediante desarrollos en serie de Taylor y bajo ciertas condiciones de regularidad sobre f, se obtiene una aproximación asintótica, supuesto que h depende de n y que $h \to 0$ y $nh \to \infty$ cuando $n \to \infty$, cuya expresión viene dada por

$$\begin{aligned} \mathrm{MISE}(\widehat{f}_h) &= \frac{1}{4}h^4\mu_2^2(K)R(f'') + \frac{1}{nh}R(K) + o\left(h^4\right) + o\left(\frac{1}{nh}\right) \\ &\equiv \mathrm{AMISE}(\widehat{f}_h) + o\left(h^4\right) + o\left(\frac{1}{nh}\right) \end{aligned} \tag{1.21}$$

siempre que los valores de R y μ_2 sean finitos, el error cuadrático medio integrado asintótico (AMISE), es la parte del MISE que es conocido de manera exacta.

El AMISE se emplea como criterio de error a optimizar en la obtención de la expresión del parámetro de suavizado óptimo. Para ello, se minimiza la expresión del AMISE en función de h despreciando los términos de menor orden y se obtiene

$$h_{\mathrm{AMISE}} = \left(\frac{R(K)}{n\mu_2^2(K)R(f'')}\right)^{1/5}.$$

Sustituyendo el valor del parámetro de suavizado óptimo en la expresión (1.21), puede comprobarse que la tasa de convergencia del error óptimo para el estimador es $o(n^{-4/5})$; ver por ejemplo Wand y Jones (1995).

Hasta ahora, se ha presentado brevemente la teoría de optimización relativa al parámetro de suavizado. También existe teoría de optimización sobre el tipo de función núcleo, se trata de escoger "la mejor" de todas las posibles funciones núcleo K, esto es, la que proporcione estimadores con mejores propiedades.

En Wand y Jones (1995) se describe dicho problema, expresando el error AMISE del estimador como producto de dos factores, de manera que uno de ellos dependa únicamente de h y el otro de K. Para esto se considera la función núcleo reescalada $K_\delta(\cdot) = (1/\delta)K(\cdot/\delta)$, pues de no hacerlo así, la expresión AMISE dada en (1.21) no se puede minimizar fácilmente en K, ya que el efecto del parámetro de suavizado y el núcleo van emparejados.

El AMISE puede factorizarse en dos términos que separan la dependencia de h y del núcleo reescalado de la forma

$$\text{AMISE}(\widehat{f_h}) = C(K_{\delta_0}) \left[\frac{1}{4} h^4 R(f'') + \frac{1}{nh} \right], \tag{1.22}$$

donde $C(K_\delta) = \left(R^4(K_\delta)\mu_2^2(K_\delta)\right)^{1/5}$ y $\delta_0 = \left(R(K)/\mu_2^2(K)\right)^{1/5}$.

Para obtener el núcleo óptimo, se debe minimizar la expresión (1.22) con respecto a K_δ, esto es, determinar la función K que minimice $C(K_{\delta_0})$.

Finalmente, tal y como indican en Wand y Jones (1995), se obtiene como núcleo óptimo el de Epanechnikov, cuya expresión es $K(x) = \dfrac{3}{4}(1 - x^2)I\left[\mid x \mid \leq 1\right]$.

Otro asunto importante a tratar sería el de encontrar un parámetro de suavizado óptimo, tema que será tratado en el Capítulo 3.

1.6.2. Estimación de la función de intensidad de un proceso de recuento

Esta sección se basa en el trabajo de Ramlau-Hansen (1983), que desarrolla estimadores de tipo núcleo para la función de intensidad de un proceso de recuento y sus derivadas. Como caso particular se obtiene el estimador de la función de azar, considerando el tiempo de supervivencia desde el enfoque de los procesos de recuento.

Sea $\{N(t);\, t \in [0, \tau]\}$ un proceso de recuento en un espacio de probabilidad $(\Omega, \mathcal{F}, P)$, con $E\left[N(\tau)\right] < \infty$ y $\{\mathcal{F}_t : t \in [0, \tau]\}$ una sub-σ-álgebra de $\mathcal{F}$, satisfaciendo las *condiciones usuales* de Andersen *et al.* (1993). Se supone que el proceso de intensidad $\lambda(t)$ asociado a $N(t)$ tiene la estructura multiplicativa de Aalen (1.2).

Bajo el marco del modelo de intensidad multiplicativo, se supone que los procesos N e Y son observables y el interés se centra en estimar α y sus derivadas. Basándose en la idea del suavizado de tipo núcleo, Ramlau-Hansen (1983) propuso el siguiente estimador constante local para α.

Sea el proceso de recuento multivariante $N = (N_1, N_2, \ldots, N_n)$, donde $N_i(t)$ tiene de intensidad a $\lambda_i(t) = \alpha(t)Y_i(t)$. El estimador de Nelson-Aalen de $A(t) = \int_0^t \alpha(s)ds$ viene dado por la expresión

$$\widehat{A}(t) = \int_0^t \frac{J^{(n)}(s)}{Y^{(n)}(s)} dN^{(n)}(s),$$

donde $J^{(n)}(s) = I\left[Y^{(n)}(s) > 0\right]$, $N^{(n)}(t) = \sum_{i=1}^n N_i(t)$, $Y^{(n)}(t) = \sum_{i=1}^n Y_i(t)$ y donde se adopta el criterio $0/0 = 0$.

Básicamente, lo que se realiza es una estimación de $\alpha(t)$ mediante un suavizado de tipo núcleo del estimador de Nelson-Aalen $\widehat{A}(t)$.

Sea K una función tipo núcleo, acotada y que se anula fuera del intervalo $[-1, 1]$ y sea h un parámetro de suavizado positivo. El estimador de $\alpha(t)$ viene dado por la expresión

$$\widehat{\alpha}(t) = \frac{1}{h}\int_0^\tau K\left(\frac{t-s}{h}\right) d\widehat{A}(s) = \frac{1}{h}\int_0^\tau K\left(\frac{t-s}{h}\right) \frac{J^{(n)}(s)}{Y^{(n)}(s)} dN^{(n)}(s).$$

Ramlau-Hansen también propuso un estimador para la derivada de α, derivando su función estimada. Por tanto, la derivada ν-ésima de α se estima por

$$\widehat{\alpha}^{(\nu)}(t) = \frac{1}{h^{\nu+1}}\int_0^\tau K^{(\nu)}\left(\frac{t-s}{h}\right) \frac{J^{(n)}(s)}{Y^{(n)}(s)} dN^{(n)}(s),$$

bajo la condición de que el núcleo K es una función continua de orden mayor o igual que ν. Los estimadores anteriores vienen definidos para $t \in [h, 1-h]$.

Ejemplo: Estimador de la función de azar de una v.a. tiempo de fallo

Sean los tiempos de fallo $T_1, T_2, \ldots, T_n$, con valores en $[0, \tau]$, con tasa de fallo α, y función de distribución $F(\tau) = 1 - \exp\left(-\int_0^\tau \alpha(s)ds\right)$ y sean $C_1, C_2, \ldots, C_n$

los correspondientes tiempos de censura i.i.d. con función de distribución G. Se supone que los tiempos de censura son independientes de los tiempos de fallo. El número de fallos

$$N^{(n)}(t) = \sum_{i=1}^{n} I\left[T_i \leq t, T_i \leq C_i\right]$$

es entonces un proceso de recuento con proceso de intensidad $\alpha(t)Y^{(n)}(t)$, donde $Y^{(n)}(t) = \sum_{i=1}^{n} I\left[T_i \geq t, C_i \geq t\right]$ indica el número de individuos vivos y en riesgo justo antes del instante t. Como se ha visto antes, $J^{(n)}(s) = I\left[Y^{(n)}(s) > 0\right]$, el estimador de la función de azar acumulado $A(t) = \int_0^t \alpha(s)ds$ es el estimador de Nelson-Aalen

$$\widehat{A}(t) = \int_0^t \frac{J^{(n)}(s)}{Y^{(n)}(s)} dN^{(n)}(s) = \sum_{T_j \leq t} \frac{\delta_j}{Y^{(n)}(T_j)},$$

donde δ_j es el indicador de fallo (muerte) para el individuo j-ésimo y donde al igual que antes se adopta el criterio $0/0 = 0$. El correspondiente estimador tipo núcleo es

$$\widehat{\alpha}(t) = \frac{1}{h}\int_0^\tau K\left(\frac{t-s}{h}\right) d\widehat{A}(s) = \frac{1}{h}\sum_{j=1}^{n} K\left(\frac{t-T_j}{h}\right)\frac{\delta_j}{Y^{(n)}(T_j)}.$$

Propiedades asintóticas del estimador de Ramlau-Hansen

El error de estimación estocástico se descompone de la siguiente forma

$$\widehat{\alpha}(t) - \alpha(t) = (\widehat{\alpha}(t) - \widetilde{\alpha}(t)) + (\widetilde{\alpha}(t) - \alpha(t))$$

donde

$$\widetilde{\alpha}(t) = \frac{1}{h}\int_0^\tau K\left(\frac{t-s}{h}\right) dA(s),$$

de modo que el error se expresa como suma de un término variable $\widehat{\alpha}(t) - \widetilde{\alpha}(t)$ y otro término estable $\widetilde{\alpha}(t) - \alpha(t)$. Con esta descripción, Ramlau-Hansen probó la consistencia y normalidad asintótica que vienen resumidas en los siguientes resultados.

Suponiendo que $0 < t_0 < t < t_1 < \tau$, que α es una función continua en t y que la función γ es una función positiva y continua en t, se tiene que

$$\sup_{s\in[t_0,t_1]}\left\{\left|\frac{Y^{(n)}(s)}{n}-\gamma(s)\right|\right\}\xrightarrow[n\to\infty]{\mathcal{P}}0.$$

Entonces, $h\to 0$ y $nh\to\infty$ cuando $n\to\infty$, por tanto

$$\sqrt{nh}(\widehat{\alpha}(t)-\widetilde{\alpha}(t))\xrightarrow[n\to\infty]{\mathcal{P}}\mathcal{N}(0,\upsilon(t)),$$

donde $\upsilon(t)=\dfrac{\alpha(t)}{\gamma(t)}\displaystyle\int_{-1}^{1}K^2(s)ds=\frac{\alpha(t)}{\gamma(t)}R(K)$.

Suponiendo además que $\alpha\in C^{(2)}([0,\tau])$ se tiene que

$$(\widetilde{\alpha}(t)-\alpha(t))\xrightarrow[n\to\infty]{\mathcal{P}}\frac{1}{2}h^2\alpha''(t)\mu_2(K).$$

El estimador de Ramlau-Hansen puede presentar un sesgo considerable en los puntos que se encuentran en la frontera. Este "efecto frontera" de los suavizadores de tipo núcleo es bastante común y debe ser tratado con especial atención en diversas situaciones. Un método para reducir el sesgo en la frontera es utilizar funciones núcleo especiales que permitan corregir en cierta medida este efecto, (ver por ejemplo Gasser y Müller (1979) o Muller y Wang (1994)). Una discusión relativa a las funciones núcleo que se adaptan bien en la frontera para el estimador de Ramlau-Hansen puede verse en el ejemplo IV.2.5 de Andersen *et al.* (1993). En Chen, Huggins, Yip y Lam (2008) y Chen, Yip y Lam (2011) se propone un estimador de la intensidad polinomial local con propiedades teóricas superiores al estimador de Ramlau-Hansen y que corrige automáticamente el efecto frontera. De forma similar, el suavizado polinomial local bajo el contexto de la regresión puede verse en Ruppert y Wand (1994), y Fan y Gijbels (1996).

Estimador de la función de intensidad de un PPNH

Bajo un modelo no paramétrico la estimación de la función de intensidad se realiza a partir de la información derivada de considerar una o múltiples realizaciones de un PPNH mediante la observación de un único sistema o de un conjunto

de sistemas con las mismas características. No se consideran hipótesis para la forma funcional de la función de intensidad $\lambda(t)$, excepto que es una función suave, suavidad que se define en términos de algunas propiedades, como la derivabilidad de la función con respecto a t.

La información muestral consiste en observaciones de tiempos de eventos de un sistema, donde los eventos ocurren en intervalos separados e independientes entre sí. Se supone que no suceden eventos simultáneos. Se observa el sistema durante un periodo de tiempo $[0, \tau]$ en el que el número total de ocurrencias se han registrado en los tiempos $T_1 < T_2 < \ldots < T_n = \tau$. Por tanto, la información muestreada consiste en una realización de un PPNH y el objetivo es encontrar un estimador no paramétrico de la función de intensidad $\lambda(t)$.

Como se describió antes, la estimación de tipo núcleo de la función de intensidad de un proceso de recuento fue introducida por Ramlau-Hansen (1983) y sus propiedades asintóticas pueden verse en Leadbetter y Wold (1983) y Diggle y Marron (1988). En particular, Ramlau-Hansen (1983) estudió la consistencia y la normalidad asintótica del estimador de la función de intensidad de un proceso de recuento que se ajusta a un modelo multiplicativo de Aalen. Se considera al PPNH como un caso particular en el que las funciones de intensidad condicionada y no condicionada coinciden, bajo el modelo multiplicativo, como se describió en (1.16).

En el caso en el que los datos están regularmente espaciados a lo largo del tiempo, puede construirse un estimador de la función de intensidad utilizando funciones de tipo núcleo de la siguiente forma

$$\widehat{\lambda}_h(t) = \frac{1}{h} \sum_{i=1}^{n} K\left(\frac{t - T_i}{h}\right), \tag{1.23}$$

donde K es una función núcleo, habitualmente se supone que es una densidad simétrica, no negativa y acotada, y h es el parámetro de suavizado que cuantifica el nivel de suavizado a considerar. Cabe destacar que este estimador es similar al estimador de tipo núcleo de una densidad excepto por el factor de normalización,

que en la expresión (1.23) no es necesario ya que la función λ no se espera que integre 1, sino que $\int_0^\tau \lambda(u)du = E[N(\tau)]$. Por otra parte, el tamaño de la muestra, es decir, el número total de eventos observados en el intervalo fijo $[0, \tau]$ es aleatorio.

Algunos trabajos recientes sobre la función de intensidad que consideran el estimador tipo núcleo (1.23) incluyen Cowling *et al.* (1996) y Phillips (2000, 2001).

La metodología de tipo núcleo se basa en la aproximación del valor de la función de intensidad $\lambda(t)$ en cada instante t, mediante una constante. De este modo las estimaciones se calculan localmente en un entorno definido por el parámetro de suavizado h. El núcleo K asigna la ponderación a cada tiempo de supervivencia observado dentro del entorno considerado. Cuando los eventos ocurren regularmente en el intervalo $[0, \tau]$, el estimador de tipo núcleo proporciona una buena aproximación de la función de intensidad.

1.6.3. Estimación de la función de supervivencia

En esta sección se considera el estimador suavizado para la función de supervivencia propuesto por Kulasekera *et al.* (2001). Se consideran T una v.a. correspondiente a los tiempos de vida bajo estudio y $T_1, T_2, \ldots, T_n = \tau$ tiempos de vida positivos i.i.d. de la función de distribución $F(t)$ con función de supervivencia $S(t) = 1 - F(t)$. Además $C_1, C_2, \ldots, C_n$ son v.a. de censura, positivas i.i.d. de una distribución $G(t)$. Bajo el modelo de censura aleatoria por la derecha, T_i está censurada por la derecha por C_i, por tanto sólo se observan las parejas (Z_i, δ_i) , $i = 1, 2, \ldots, n$, siendo Z_i el mínimo de T_i y C_i y δ_i el indicador del evento $T_i \leq C_i$. Esto es, para $i = 1, 2, \ldots, n$ se tiene

$$Z_i = T_i \wedge C_i, \quad \delta_i = \begin{cases} 1 & \text{si } T_i \leq C_i, \\ 0 & \text{en otro caso.} \end{cases}$$

Por tanto, Z_i, $i = 1, 2, \ldots, n$, son variables aleatorias i.i.d. con función de distribución H dada por

$$1 - H(s) = (1 - F(s))(1 - G(s)), \;\; 0 < s < \infty.$$

El estimador de la función de supervivencia propuesto por Breslow (1972) viene dado por

$$\widehat{S}_B(t) = \exp\left(-\sum_{Z_{(i)}\leq t} \frac{\delta_{(i)}}{Y(Z_{(i)})}\right) \tag{1.24}$$

donde $Y(Z_{(i)})$ es el número de individuos en riesgo (es decir, vivos y no censurados) justo antes del instante $Z_{(i)}$.

A continuación, se define el estimador de la función de supervivencia de Kulasekera *et al.* (2001). Se considera K una función núcleo que satisface las propiedades de las funciones núcleo anteriormente citadas. El estimador suavizado de la función de supervivencia viene dado por la función $\widetilde{S}$ y se define

$$\widetilde{S}(t) = \frac{1}{h}\int_0^\tau K\left(\frac{t-s}{h}\right)\widehat{S}_B(s)ds, \;\; 0 < t \leq \tau$$

donde h es el parámetro de suavizado y $\widehat{S}_B(s)$ es el estimador de Breslow definido en (1.24).

Propiedades asintóticas del estimador

Se supone que la función S es continuamente diferenciable satisfaciendo la condición de Hölder

$$|S(t) - S(s)| \leq \zeta|t-s|^\vartheta, \;\; \forall \zeta, \vartheta > 0.$$

Además, se supone también que $T_H = \inf\{t : H(t) = 1\}$ para cualquier función de distribución H y $T^* < T_H$. Entonces cuando $n \to \infty$ el sesgo, $E[\widehat{S}] - S$, converge exponencialmente a cero en $[0, T^*]$. Además, se tiene que $\widehat{S}$ es fuertemente consistente (ver Aalen (1978)) y que $\sqrt{n}(\widehat{S} - S)$ converge débilmente a un proceso Gaussiano de media cero sobre el espacio de Skorokhod $D([0,t])$. Por lo tanto, Kulasekera *et al.* (2001) proporciona la consistencia uniforme fuerte del estimador $\widehat{S}(t)$ en $[0, T_H]$.

$$\sup_{0\leq t\leq T_H}\left\{\left|(\widetilde{S}(t) - S(t))\right|\right\} \xrightarrow[n\to\infty]{c.s.} 0.$$

cuya tasa de convergencia es $o\left(1/\sqrt{n}\right)$, que sigue un argumento similar al de Fernholz (1991).

1.6.4. Estimadores locales mínimo-cuadráticos

El uso de métodos de estimación polinomial local se ha popularizado en el contexto de la regresión no paramétrica, ver por ejemplo, Wand y Jones (1995) o Fan y Gijbels (1996). En el caso de la función de densidad es aún hoy en día menos conocido. No obstante Jones (1993) propuso un estimador lineal local para la densidad que fue extendido más tarde para la estimación de la función de azar. En concreto Nielsen y Tanggaard (2001) presentan una clase de estimadores lineales locales para funciones de azar unidimensionales, basados en una estimación de mínimos cuadrados ponderados con diferentes esquemas de ponderación. En dicho trabajo, se muestra que el estimador de Ramlau-Hansen (1983) se relaciona de forma particular con el estimador lineal local. Si el estimador de Ramlau-Hansen se corrige por la frontera con un conjunto particular de funciones núcleos definidos en Gasser y Müller (1979), entonces se puede interpretar como un estimador lineal local de la función de azar con un tipo en particular de ponderación que Nielsen y Tanggaard (2001) llaman ponderación de Ramlau-Hansen. Este estimador se describe a continuación con detalle bajo la formulación siguiente.

Se observan n individuos y se define N_i como los recuentos de los fallos observados para el i-ésimo individuo en el intervalo $[0, \tau]$. Se supone que N_i es un proceso de recuento unidimensional no decreciente, continuo a la derecha adaptado a la filtración $\{\mathcal{F}_t : t \in [0, \tau]\}$, es decir, que obedece las *condiciones usuales* de Andersen *et al.* (1993). Se considera nuevamente que la función de intensidad λ_i sigue el modelo multiplicativo de Aalen, $\lambda_i(t) = \alpha(t)Y_i(t)$, sin restricciones en la forma funcional de $\alpha(\cdot)$, donde Y_i es un proceso predecible que toma el valor 1 cuando el i-ésimo individuo está en riesgo y bajo observación, y el valor 0 en otro caso. Se supone que $(N_1, Y_1), (N_2, Y_2), \ldots, (N_n, Y_n)$ son i.i.d. para los n individuos. Todas las

martingalas y procesos predecibles se definen con respecto a la filtración $\mathcal{F}_t$. Con estas definiciones, λ_i es predecible y el proceso $M_i(y) = N_i(y) - \Lambda_i(y)$, $i = 1, 2, \ldots, n$ con $\Lambda_i(y) = \int_0^y \lambda_i(s)ds$ son martingalas locales de cuadrado integrable. Se define también $Y^{(n)}(s) = \sum_{i=1}^n Y_i(s)$ y $J^{(n)}(s) = I\left[Y^{(n)}(s) > 0\right]$.

Estimador de la función de azar

Sea K una función de densidad de probabilidad con soporte $[-1, 1]$, simétrica alrededor del cero y sea $K_h(\cdot) = h^{-1}K(\cdot/h)$ para cualquier parámetro de suavizado h. Sea $W(s)$ una función peso arbitraria y sea $q_p(z) = \sum_{i=0}^p \theta_i z^i$ una función polinómica de grado p. A continuación se define un estimador constante local de tipo núcleo de la función de azar basado en el principio de mínimos cuadrados introducido por Nielsen (1998a). Dicho principio de mínimos cuadrados está íntimamente relacionado al principio introducido para la estimación de la función de densidad en Jones (1993). Se define

$$\widehat{\theta}(t) = \underset{\theta}{\operatorname{argmin}} \left\{ \sum_{i=1}^n \int_0^\tau \left[\Delta N_i(s) - q_p(t-s)\right]^2 K_h(t-s)W(s)Y_i(s)ds \right\}. \quad (1.25)$$

El estimador constante local surge cuando el grado del polinomio es $p = 0$, esto es, $q_0(t - s) = \theta_0$, por lo que, resolviendo la expresión (1.25) con respecto a θ_0, y despejando esta se obtiene

$$\widehat{\theta}_0 = \widehat{\alpha}_{0,W}(t) = \frac{\displaystyle\sum_{i=1}^n \int_0^\tau K_h(t-s)W(s)Y_i(s)dN_i(s)}{\displaystyle\sum_{i=1}^n \int_0^\tau K_h(t-s)W(s)Y_i(s)ds}$$

donde $\widehat{\alpha}_{0,W}(t)$ es el estimador constante local de $\alpha(t)$ basado en la función peso W.

El estimador lineal local se obtiene considerando $p = 1$, esto es, $q_1(t - s) = \theta_0 + \theta_1(t - s)$, por lo que, resolviendo la expresión (1.25) con respecto a θ_0 y θ_1, y

despejando θ_0 se obtiene

$$\widehat{\theta}_0 = \widehat{\alpha}_{1,W}(t) = \sum_{i=1}^{n} \int_0^\tau \overline{K}_{t,h}(t-s)dN_i(s),$$

donde

$$\overline{K}_{t,h}(t-s) = \frac{a_2(t) - a_1(t)(t-s)}{a_0(t)a_2(t) - a_1^2(t)} K_h(t-s)W(s),$$

con

$$a_j(t) = \int_0^\tau K_h(t-s)(t-s)^j W(s)Y^{(n)}(s)ds \quad \text{para } \ j = 0, 1, 2, \tag{1.26}$$

además, se cumple que

$$\int_0^\tau \overline{K}_{t,h}(t-s)Y^{(n)}(s)ds = 1, \quad \int_0^\tau \overline{K}_{t,h}(t-s)(t-s)Y^{(n)}(s)ds = 0,$$
$$\int_0^\tau \overline{K}_{t,h}(t-s)(t-s)^2 Y^{(n)}(s)ds > 0.$$

Los estimadores constante local y lineal local construidos con la función de pesos de Ramlau-Hansen $W(s) = nJ^{(n)}(s)/Y^{(n)}(s)$ son iguales al estimador de Ramlau-Hansen en el interior del intervalo de estimación. El estimador lineal local con la función de pesos unitarios $W(s) = 1$ es un estimador más robusto que el estimador lineal local con la función de pesos de Ramlau-Hansen. Los pesos unitarios son también los análogos más cercanos de los estimadores lineales locales con el enfoque tradicional en regresión lineal local; ver Fan y Gijbels (1996).

Propiedades asintóticas del estimador lineal local de la función de azar

Las demostraciones de las propiedades asintóticas del estimador de la función de azar siguen la teoría asintótica estándar de las martingalas (ver Ramlau-Hansen (1983)), esto puede encontrarse en Nielsen (1998a) o Nielsen y Tanggaard (2001). En dichos trabajos se muestra que todos los estimadores lineales locales tienen la misma distribución asintótica independientemente de la función peso W que se utilice.

En primer lugar, se consideran las siguientes hipótesis generales:

(A.1) El parámetro de suavizado $h \to 0$ *y* $nh \to \infty$*, cuando* $n \to \infty$*.*

(A.2) Existen las funciones positivas γ *y* w *tal que*

$$\sup_{s\in[0,\tau]}\left\{\left|\frac{Y^{(n)}(s)}{n}-\gamma(s)\right|\right\}\xrightarrow[n\to\infty]{\mathcal{P}}0$$

y

$$\sup_{s\in[0,\tau]}\left\{\left|\frac{W(s)}{n}-w(s)\right|\right\}\xrightarrow[n\to\infty]{\mathcal{P}}0,$$

(A.3) γ, $w \in C^{(1)}([0,\tau])$.

(A.4) $\alpha \in C^{(2)}([0,\tau])$.

□

El error de estimación estocástico se descompone de la siguiente forma

$$\widehat{\alpha}_W(t)-\alpha(t)=(\widehat{\alpha}_W(t)-\widetilde{\alpha}_W(t))+(\widetilde{\alpha}_W(t)-\alpha(t))$$

donde

$$\widetilde{\alpha}_W(t)=\sum_{i=1}^{n}\int_0^\tau \overline{K}_{t,h}(t-s)d\Lambda_i(s).$$

de modo que el error se expresa como suma de un término variable $V_W(t)=\widehat{\alpha}_W(t)-\widetilde{\alpha}_W(t)$ y un término estable $B_W(t)=\widetilde{\alpha}_W(t)-\alpha(t)$.

El término del sesgo puede aproximarse mediante un desarrollo de Taylor obteniéndose

$$B_W(t)=\widetilde{\alpha}_W(t)-\alpha(t)=\frac{1}{2}h^2\alpha''(t)\mu_2(K)+o_P(h^2),$$

y el término de la varianza es

$$V_W(t)=\widehat{\alpha}_W(t)-\widetilde{\alpha}_W(t)=\sum_{i=1}^{n}\int_0^\tau \overline{K}_{t,h}(t-s)dM_i(s)$$

que es asintóticamente equivalente a

$$\sum_{i=1}^{n}\int_0^\tau K_h(t-s)\frac{dM_i(s)}{Y^{(n)}(s)}.$$

Suponiendo que las hipótesis generales se mantienen, entonces

$$\sqrt{nh}V_W(t) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \upsilon_W(t)\right)$$

siendo

$$\upsilon_W(t) = \frac{\alpha(t)}{\gamma(t)}R(K)$$

y

$$B_W(t) = h^2 m_W(t) + o_P(h^2),$$

donde para el estimador constante local y según sea la función peso de Ramlau-Hansen o natural, la expresión de $m_W(t)$ será

$$m_{\text{R-H}}(t) = \left(\frac{\alpha''(t)}{2}\right)\mu_2(K), \qquad m_N(t) = \left(\frac{\alpha''(t)}{2} + \frac{\alpha'(t)\gamma'(t)}{\gamma(t)}\right)\mu_2(K)$$

y para el estimador lineal local con la función peso natural, la expresión de $m_W(t)$ será

$$m_N(t) = \left(\frac{\alpha''(t)}{2}\right)\mu_2(K).$$

Mientras que los estimadores constantes locales construidos mediante los dos tipos de pesos considerados tienen las mismas varianzas asintóticas, existe una diferencia con respecto al sesgo. El sesgo asociado con el estimador constante local utilizando los pesos naturales, tiene un término más que el sesgo asociado al estimador que utiliza los pesos de Ramlau-Hansen. Este término extra hace que sean preferibles los pesos de Ramlau-Hansen en este caso en particular. Es fácil observar que ninguno de estos estimadores son asintóticamente mejores para cada combinación de α y γ. La teoría asintótica sólo se mantiene en el interior del intervalo, cerca de las fronteras el sesgo de estos estimadores constantes tiene una tasa de convergencia inferior; ver Jones (1993).

A diferencia del estimador constante local donde las propiedades asintóticas dependen de los pesos considerados, la situación es algo diferente para el caso lineal local. Estos estimadores tienen un comportamiento puntual que es independiente

de la ponderación elegida. En Nielsen y Tanggaard (2001) se muestra mediante un estudio de simulación que la ponderación unitaria mejora considerablemente a la ponderación de Ramlau-Hansen, por lo que se concluye que el peso unitario es más robusto a la hora de detectar patrones complejos que el de Ramlau-Hansen.

Estimador lineal local de la función de densidad

Considerando las mismas condiciones que en la sección anterior para estimar la función de azar, en esta sección se procede de manera análoga para calcular un estimador de la función de densidad. Como es sabido, la relación existente entre las funciones de densidad, riesgo y supervivencia viene dada por la relación $f = \alpha \cdot S$, por tanto, puede definirse un estimador polinomial local de la densidad de tipo núcleo, basándose en una aproximación de mínimos cuadrados de Nielsen (1998b). Se tiene también que $\widehat{f}_{p,W}(t) = \widehat{\theta}_0(t) = \widehat{\theta}_0$, donde $\widehat{\theta} = \left(\widehat{\theta}_0, \widehat{\theta}_1, \ldots, \widehat{\theta}_p\right)$, por lo que, operando de manera análoga a la estimación del riesgo se llega al siguiente problema de mínimos cuadrados

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \int_0^\tau \left[\widehat{S}(s)\Delta N_i(s) - q_p\left(t-s\right) \right]^2 K_h(t-s)W(s)Y_i(s)ds \right\}.$$

Al igual que en el apartado anterior, haciendo $p = 0$ para el estimador constante local, se tiene

$$\widehat{f}_{0,W}(t) = \frac{\displaystyle\sum_{i=1}^{n} \int_0^\tau K_h(t-s)W(s)Y_i(s)\widehat{S}(s)dN_i(s)}{\displaystyle\int_0^\tau K_h(t-s)W(s)Y^{(n)}(s)ds},$$

y para $p = 1$ se llega al estimador lineal local

$$\widehat{f}_{1,W}(t) = \sum_{i=1}^{n} \int_0^\tau \overline{K}_{t,h}(t-s)W(s)Y_i(s)\widehat{S}(s)dN_i(s),$$

donde $\overline{K}_{t,h}(t-s)$ se definió en el apartado anterior.

Los estimadores de la densidad que se tratan en esta sección implican a un estimador de la función de supervivencia, que generalmente se escribe $\widehat{S}(t)$. Para estos datos de tiempos de vida puede considerarse, por ejemplo, el estimador producto límite de la función de supervivencia de Kaplan-Meier; ver Fleming y Harrington (1991).

Para cualquier función de peso W dada, en Nielsen *et al.* (2009) se demuestra que el estimador lineal local es más conveniente que el estimador constante local ya que la teoría y las simulaciones así lo demuestran. Las propiedades asintóticas del sesgo del estimador constante local son menos atractivas que las del estimador lineal local y además, en presencia de fronteras conocidas, el estimador lineal local proporciona una buena corrección de las fronteras con respecto al estimador constante local.

Propiedades asintóticas del estimador lineal local de la densidad

Las propiedades asintóticas del estimador lineal local de la función de densidad siguen un paralelismo con las propiedades asintóticas del estimador lineal local para la función de azar.

Teorema 3. *Suponiendo que las hipótesis generales se mantienen y que* $f \in C^{(2)}([0,\tau])$*, entonces*

$$\sqrt{nh}V_W(t) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \upsilon_W(t)\right)$$

y

$$B_W(t) = h^2 m_W(t) + o_P(h^2),$$

donde

$$m_N(t) = \left(\frac{f''(t)}{2}\right)\mu_2(K),$$

siendo

$$\upsilon_W(t) = \frac{f(t)S(t)}{\gamma(t)}R(K).$$

Dichas propiedades, además de un método de selección del parámetro de suavizado óptimo basado en validación cruzada, pueden encontrarse en Nielsen *et al.* (2009). En Gámiz-Pérez, Martínez-Miranda y Nielsen (2013a) se desarrolla un novedoso método para seleccionar el parámetro de suavizado óptimo. Este método es más sencillo y mejora al método clásico plug-in para este tipo de estimadores de la densidad.

1.6.5. El parámetro de suavizado en la estimación de la función de intensidad

El estimador tipo núcleo de la función de intensidad propuesto en (1.23) involucra a un parámetro de suavizado desconocido h que debe estimarse. A continuación se describen dos de los métodos más populares para la selección del parámetro de suavizado, como son el método de validación cruzada y el método plug-in.

Método de validación cruzada

El método de validación cruzada propuesto por Rudemo (1982) y Bowman (1984) es un procedimiento completamente automático para la selección del parámetro de suavizado. El método de validación cruzada ha sido ampliamente utilizado en otros contextos de estimación tipo núcleo, como en la estimación de la densidad, ver Silverman (1986). En esta sección, se hace referencia al análisis de Brooks y Marron (1991), donde se adaptan los argumentos utilizados en la estimación tipo núcleo de la densidad para obtener los de la función de intensidad. Este razonamiento se basa en el trabajo de Ramlau-Hansen (1983), donde los autores adoptan el error cuadrático integrado (ISE) como una medida de la estimación del error, que viene definido en el contexto de las funciones de intensidad de la forma

$$\begin{aligned}\mathrm{ISE}(h) = \mathrm{ISE}(\widehat{\lambda}_h(t)) &= \int_0^\tau \left(\widehat{\lambda}_h(t) - \lambda(t)\right)^2 dt \\ &= \int_0^\tau \widehat{\lambda}_h^2(t)dt - 2\int_0^\tau \widehat{\lambda}_h(t)\lambda(t)dt + \int_0^\tau \lambda^2(t)dt,\end{aligned}$$

sabiendo que $\widehat{\lambda}_h$ viene definida en (1.23), y que el último término no depende de

h. Por lo tanto, una elección ideal del parámetro de suavizado (en el sentido de minimizar el ISE) consiste en minimizar los dos primeros términos

$$\underset{h}{\operatorname{argmin}} \left\{ \int_0^\tau \widehat{\lambda}_h^2(t)dt - 2\int_0^\tau \widehat{\lambda}_h(t)\lambda(t)dt \right\}. \tag{1.27}$$

El principio básico del método de validación cruzada es obtener una estimación del segundo término de (1.27), sustituirlo y minimizar la expresión resultante con respecto a h. Con esta idea se define $\widehat{\lambda}_{-i}$ como un estimador de la intensidad construido a partir de todos los valores del conjunto de datos excepto T_i, es decir, un estimador "leave-one-out" definido de la forma

$$\widehat{\lambda}_{-i}(t) = \frac{1}{h}\sum_{j\neq i} K\left(\frac{t-T_j}{h}\right).$$

Finalmente, se define la función de validación cruzada sugerida por Brooks y Marron (1991),

$$\mathrm{CV}(h) = \int_0^\tau \widehat{\lambda}_h^2(t)dt - 2\sum_{i=1}^n \widehat{\lambda}_{-i}(T_i) \tag{1.28}$$

que sólo depende de los datos.

Además, puede comprobarse que

$$E\left[\sum_{i=1}^n \widehat{\lambda}_{-i}(T_i)\right] = E\left[\int_0^\tau \widehat{\lambda}(t)\lambda(t)dt\right], \tag{1.29}$$

y sustituyendo (1.29) en la expresión (1.28) se demuestra que $\mathrm{CV}(h) + \int_0^\tau \lambda^2(t)dt$ es una buena aproximación del error cuadrático medio integrado (MISE)

$$\mathrm{MISE}(h) = \mathrm{MISE}(\widehat{\lambda}_h(t)) = E\left[\int_0^\tau \left(\widehat{\lambda}_h(t) - \lambda(t)\right)^2 dt\right]. \tag{1.30}$$

Finalmente, se supone que la minimización de $E[\mathrm{CV}(h)]$ es un valor muy próximo al que proporciona la minimización de $\mathrm{CV}(h)$, por lo que se espera que este último proporcione una buena opción para el parámetro de suavizado.

Con el propósito de agilizar los cálculos computacionales de $\mathrm{CV}(h)$, se decide seguir un procedimiento análogo al explicado en Silverman (1986) en el contexto de la estimación de la densidad.

Si se supone que la función núcleo K es simétrica, realizando el cambio de variable $u = t/h$ en la integral de la ecuación (1.28), se tiene que

$$\begin{aligned}\int_0^\tau \widehat{\lambda}_h^2(t)dt &= \int_0^\tau \sum_{i=1}^n h^{-1}K\left(\frac{t-T_i}{h}\right) \times \sum_{j=1}^n h^{-1}K\left(\frac{t-T_j}{h}\right) dt \\ &= h^{-1}\sum_{i=1}^n\sum_{j=1}^n \int_0^\tau K(h^{-1}T_i - u)K(u - h^{-1}T_j)du \\ &= h^{-1}\sum_{i=1}^n\sum_{j=1}^n K^{(2)}\left(\frac{T_i - T_j}{h}\right),\end{aligned}$$

donde $K^{(2)}$ es la convolución del núcleo.
El segundo término de la expresión (1.28) se obtiene como

$$\begin{aligned}\sum_{i=1}^n \widehat{\lambda}_{-i}(T_i) &= \sum_{i=1}^n \left\{\sum_{j=1}^n h^{-1}K\left(\frac{T_i - T_j}{h}\right) - h^{-1}K(0)\right\} \\ &= \sum_{i=1}^n\sum_{j=1}^n h^{-1}K\left(\frac{T_i - T_j}{h}\right) - nh^{-1}K(0).\end{aligned}$$

Finalmente

$$\mathrm{CV}(h) = \frac{1}{h}\sum_{i=1}^n\sum_{j=1}^n \left\{K^{(2)}\left(\frac{T_i - T_j}{h}\right) - 2K\left(\frac{T_i - T_j}{h}\right)\right\} + \frac{2n}{h}K(0).$$

Método plug-in

Otro criterio utilizado para evaluar la precisión global del estimador $\widehat{\lambda}_h$ es el MISE dado por la expresión (1.30). Se han propuesto varios selectores del parámetro de suavizado basados en el criterio MISE. Se considera la aproximación del MISE definida a partir de los términos del sesgo y la varianza

$$\begin{aligned}\mathrm{MISE}(h) = \mathrm{MISE}(\widehat{\lambda}_h(t)) &= \int_0^\tau \left(E[\widehat{\lambda}_h(t)] - \lambda(t)\right)^2 dt + \int_0^\tau \mathrm{Var}[\widehat{\lambda}_h(t)]dt \\ &= \frac{1}{4}h^4\mu_2^2(K)\int_0^\tau (\lambda''(t))^2 dt + \frac{1}{h}\int_0^\tau K^2(t)dt + o\left(h^4\right) + o\left(\frac{1}{h}\right),\end{aligned} \tag{1.31}$$

donde $\mu_2(K) = \int x^2K^2(x)dx$, (para más detalles ver Silverman (1986)). Dada la expresión (1.31), se puede definir el error cuadrático medio integrado asintótico

(AMISE) de la forma

$$\text{AMISE}(h) = \text{AMISE}(\widehat{\lambda}_h(t)) = \frac{1}{4}h^4\mu_2^2(K)R(\lambda'') + \frac{1}{h}R(K)$$

donde $R(g) = \int g^2(x)dx$ y despreciándose los términos de menor orden. Mediante un cálculo sencillo, el parámetro de suavizado óptimo que minimiza el AMISE viene dado por

$$h_{\text{AMISE}} = \underset{h}{\text{argmin}}\ \{\text{AMISE}(h)\} = \left(\frac{R(K)}{\mu_2^2(K)R(\lambda'')}\right)^{1/5}$$

Sin embargo, $R(\lambda'')$ es desconocida por lo que en la práctica, no se puede calcular el parámetro de suavizado óptimo. Una forma de obtener un estimador de dicho valor óptimo consiste en asumir una forma conocida de la función λ. La conocida regla del pulgar ("rule-of-thumb") se basa en esta idea y fue introducida por Silverman (1986) en el contexto de la estimación de la densidad.

Ejemplo: Selectores del parámetro de suavizado

A continuación se ilustran los selectores del parámetro de suavizado antes descritos en un ejemplo con datos de Fiabilidad. Se consideran los 136 tiempos entre fallos del System Code 1 que pueden obtenerse via internet en The Data & Analysis Center for Software (DACS). Esto es parte de unos datos referidos a tiempos de fallo del software en 16 proyectos y que fueron compilados por John Musa de Bell Telephone Laboratories. El conjunto de datos tratado consiste en tiempos de fallo entre llegadas (en segundos) que representan el tiempo transcurrido entre fallos consecutivos.

Se supone que los fallos ocurren de acuerdo a un PPNH y el objetivo es estimar la función de intensidad de forma no paramétrica. Para ello se construye un estimador como el de la expresión (1.23) y se obtienen los parámetros de suavizado mediante los dos métodos de selección descritos anteriormente. En la Figura 1.4 pueden apreciarse las intensidades estimadas correspondientes a los datos del presente ejemplo, obtenidas con parámetros de suavizado calculados por validación cruzada y por el método plug-in. En dicha figura, se observa que existe una gran

diferencia entre las intensidades obtenidas por ambos métodos. Mientras que validación cruzada ofrece una estimación de la intensidad muy ruidosa, con muchos cambios en la tendencia, el método plug-in da lugar a una estimación sobresuavizada donde lo único que se aprecia es la tendencia global decreciente. En la siguiente sección se describirá una aproximación inferencial acerca de la función de intensidad que elimina el problema de seleccionar el parámetro de suavizado óptimo.

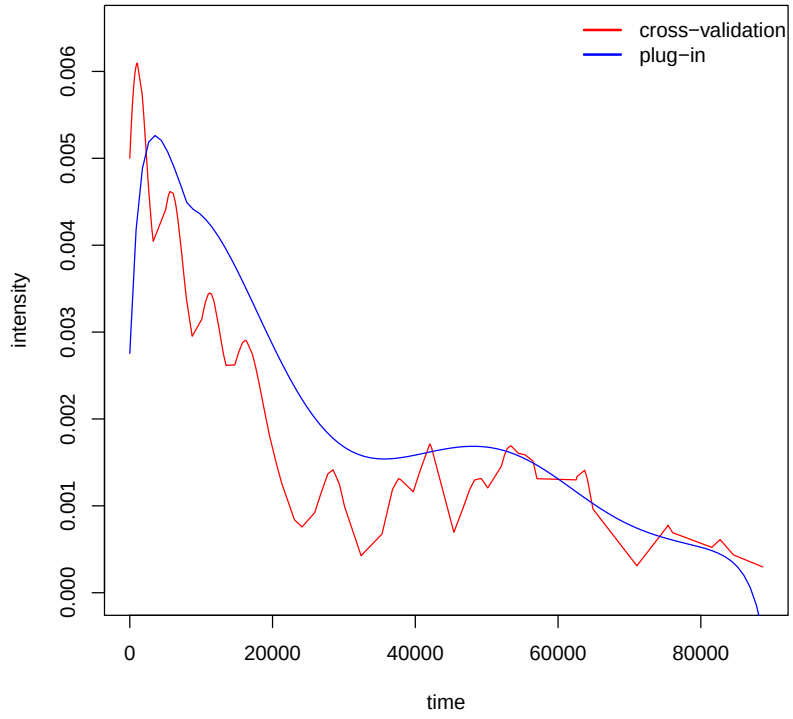


Figura 1.4: Funciones de intensidad estimadas con parámetros de suavizado obtenidos por el método de validación cruzada y por el método plug-in. Datos de fallo de software.

□

Capítulo 2

Modelos de regresión con datos de supervivencia basados en procesos de recuento

2.1. Introducción y objetivos

En el análisis de tiempos de vida la descripción del proceso de deterioro de un sistema puede requerir la consideración de determinados factores (exógenos y endógenos) a los que usualmente son referidos como *covariables* o *variables explicativas*. La inclusión de este tipo de información en el modelo que describe el deterioro del sistema ha motivado la formulación de modelos de regresión para el análisis de tiempos de vida, considerando técnicas específicas para el tratamiento de las características particulares que presentan los tiempos de vida (censura y/o truncamiento).

El modelo de riesgos proporcionales (PH) de Cox (1972) y el modelo de tiempo de vida acelerada (AFT) de Lawless (1982) y Nelson (1990), son los modelos utilizados más frecuentemente en análisis de Fiabilidad y Supervivencia.

El modelo de Cox busca la relación entre los riesgos de fallo (muerte) de dos dispositivos (individuos) expuestos a factores de riesgo diferentes. Para ello, el modelo parte de una hipótesis fundamental, la de que los riesgos son proporcionales.

La principal ventaja de este modelo es que para evaluar el efecto de las covariables sobre la distribución del tiempo de vida, no se establece ninguna hipótesis sobre la forma de la función de azar base. Sin embargo, la hipótesis de riesgos proporcionales puede no cumplirse en muchos casos prácticos. En tales situaciones, el modelo AFT ha demostrado ser una alternativa conveniente en muchos casos.

El modelo AFT establece una relación directa entre el tiempo de fallo y las covariables. Usualmente, la estimación del modelo AFT se lleva a cabo suponiendo una distribución paramétrica para los tiempos de vida. A partir de ahora, este tipo de modelo será referido como modelo AFT paramétrico.

Se han propuesto numerosos métodos para la estimación del modelo AFT paramétrico, como puede verse en Lawless (1982) y Kalbfleisch y Prentice (2002). No obstante, en la mayoría de los casos, dichos modelos AFT paramétricos resultan ser muy restrictivos. Como alternativa a ellos, existen los modelos semi-paramétricos, donde no se especifica ninguna hipótesis sobre la función de distribución de supervivencia base, lo que en la práctica suele ser más conveniente. En el trabajo de Ritov (1990) se considera la estimación general para modelos de regresión lineales con datos censurados, y en los trabajos de Tsiatis (1990), Lai y Ying (1991) o Jin, Lin, Wei y Ying (2003) se proponen métodos basados en un test de rangos para estimar los parámetros de regresión con datos censurados. Además, se han estudiado diversos métodos basados en estimaciones de mínimos cuadrados, como pueden ser Miller (1976), Buckley y James (1979) o Stute (1993).

En los trabajos de Stute (1996a, 1996b) se considera un modelo AFT semiparamétrico y se introduce un procedimiento para estimar los coeficientes de regresión bajo el supuesto de censura aleatoria. En el trabajo de Gross y Lai (1996) se desarrolla un análisis de regresión con datos de supervivencia afectados por truncamiento por la izquierda y censura por la derecha. Bajo este esquema, el tiempo de vida sólo es observable a través del rango comprendido entre el límite inferior del soporte del truncamiento y el límite superior del soporte de la censura. Los

estimadores de los parámetros de regresión son relativamente fáciles de obtener y resultan útiles para explorar la relación entre la variable respuesta y el vector de covariables.

En todos los casos anteriores, el estudio se centra en la estimación de los parámetros de regresión. Por el contrario, el interés de este capítulo va más allá de la estimación de los parámetros de regresión. La motivación práctica es evaluar el funcionamiento de un sistema de suministro de agua situado en una ciudad del litoral mediterráneo español, para lo cual se formula un modelo de regresión de tiempo de vida acelerada no paramétrico. En este caso, se considera como unidades estadísticas tramos de tubería y el problema es evaluar la probabilidad de que una tubería esté en funcionamiento transcurrido un determinado periodo de tiempo, para lo cual es preciso estimar el modelo completo.

Con este fin, se propone el uso de técnicas no paramétricas que permiten evaluar el riesgo de fallo en el sistema de suministro de agua. El término "no paramétrico" se refiere a que no se considera ninguna forma particular de la función de distribución del tiempo de vida de un individuo de la población. En otras palabras, se considera un modelo AFT semi-paramétrico que relaciona directamente el tiempo de fallo de una tubería con las características particulares que la describen (propiedades físicas de la tubería como pueden ser dimensiones o material del que está construida así como las características del entorno en el que está instalada).

Se propone un procedimiento de estimación del modelo secuencial. En primer lugar, se consideran los métodos sugeridos por Gross y Lai (1996) y Stute (1996a, 1996b) para la estimación de los parámetros de regresión. Estas estimaciones se utilizan para transformar los datos en una escala de tiempos base. Seguidamente, se lleva a cabo un procedimiento no paramétrico para estimar la función de supervivencia base. Finalmente, se realiza una transformación que proporciona el estimador de la función de supervivencia para un sujeto específico.

La función de supervivencia base se estima usando un estimador lineal local.

El estimador está íntimamente relacionado con el estimador de la función de azar propuesto por Nielsen y Tanggaard (2001) y con el estimador de la función de densidad de Nielsen *et al.* (2009).

La estructura de este capítulo es similar a la del trabajo publicado en López-Montoya, Gámiz-Pérez y Martínez-Miranda (2015), si bien, en este capítulo se proporcionan más detalles y se hace un análisis más profundo del ejemplo práctico y las simulaciones. Concretamente, en la Sección 2.2 se hace una revisión de los modelos de regresión más comunes en Fiabilidad y Supervivencia. Seguidamente, en la Sección 2.3 se presenta el modelo de tiempo de vida acelerada en términos de procesos de recuento, también se propone una estimación secuencial para obtener los coeficientes de regresión así como una estimación no paramétrica de la función de supervivencia base del modelo. En la Sección 2.4 se presenta un amplio y detallado estudio de simulación que consta principalmente de dos partes: la primera consiste en probar la precisión en la estimación de los coeficientes de regresión del modelo propuesto, la segunda consiste en evaluar la precisión del estimador lineal local de la función de supervivencia base. En la Sección 2.5 se describe una aplicación a datos reales relativa a la red de suministro de agua de una ciudad de tamaño medio de la costa mediterránea española, con el objetivo de evaluar el tiempo de vida de las tuberías utilizando la aproximación propuesta. Finalmente, en la Sección 2.6, se muestran las conclusiones generales del capítulo.

2.2. Modelos de regresión basados en la función de azar

La descripción física del proceso de deterioro de un sistema puede requerir la consideración de varios factores que comúnmente se llaman covariables o variables explicativas. La introducción de este tipo de información en el modelo de deterioro puede tratarse de varias formas, lo que da lugar a diferentes modelos de regresión para datos de tiempos de vida. Existe una amplia literatura sobre modelos que

se ocupan de la relación entre el tiempo de vida y las covariables (ver por ejemplo los libros de Andersen *et al.* (1993), Therneau y Grambsch (2000), Klein y Moeschberger (2003), Kleinbaum y Klein (2005) o Martinussen y Scheike (2006)).

La estrategia más común consiste en expresar la dependencia de la información auxiliar a través de la función de azar. En otras palabras, el riesgo instantáneo de fallo de un dispositivo en particular será formulado en términos de las características particulares del mismo. Por tanto, se define la función de azar condicionada como sigue.

Definición 19 (**Función de azar condicionada).** *Sea T una v.a. que indica el tiempo de vida de un sistema o dispositivo. Sea $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^\top$ un vector de p covariables, la función de azar condicionada de T dado $\boldsymbol{x}$ se define como*

$$\alpha(t; \boldsymbol{x}) = \lim_{\Delta \to 0} \frac{Pr\{t < T \leq t + \Delta | T > t, \boldsymbol{X} = \boldsymbol{x}\}}{\Delta}. \tag{2.1}$$

Para $t > 0$ conocido y dado $\boldsymbol{x}$, la función de azar puede escribirse como el cociente de la función de densidad condicionada $f(t; \boldsymbol{x})$ entre la función de supervivencia condicionada $S(t; \boldsymbol{x}) = 1 - F(t; \boldsymbol{x})$, esto es

$$\alpha(t; \boldsymbol{x}) = \frac{f(t; \boldsymbol{x})}{S(t; \boldsymbol{x})},$$

para $S(t; \boldsymbol{x}) > 0$.

Los modelos que se verán en esta sección cuentan con la presencia, de algún modo, de información incompleta en el conjunto de datos, que implica una importante limitación en la aplicación de los modelos tradicionales en los problemas de estadística estándar. La presencia de censura por la derecha es probablemente la característica más común que presentan los datos en los estudios de Supervivencia y Fiabilidad e implica, como se ha mencionado en secciones anteriores, la observación incompleta del tiempo de vida en estudio, debido a causas diferentes a un suceso fatal final al que el sistema es susceptible, y después del cual el sistema

dejará de existir en algún sentido. A continuación se formalizan las condiciones bajo las cuales se establecerán los métodos que posteriormente se explican.

Aunque se puede trabajar de forma general en un marco de datos de supervivencia, el interés en este capítulo se centra en un tipo particular de datos de tiempos de vida en concreto con censura por la derecha y/o truncamiento por la izquierda.

Definición 20 (**Modelo de censura aleatoria por la derecha y truncamiento por la izquierda (LTRC)).** *Se considera una v.a. T que representa el tiempo de vida, C es el tiempo aleatorio de censura por la derecha y L es el tiempo de truncamiento por la izquierda. Los datos tienen la forma* $\{L_1, Z_1, \delta_1, \boldsymbol{X}_1\}, \{L_2, Z_2, \delta_2, \boldsymbol{X}_2\}, \ldots, \{L_n, Z_n, \delta_n, \boldsymbol{X}_n\}$, *donde*

- $T_1, T_2, \ldots, T_n$ *son realizaciones independientes de una v.a. de tiempo de vida* T.
- $C_1, C_2, \ldots, C_n$ *son los valores de una v.a.* C *de censura por la derecha.*
- $L_1, L_2, \ldots, L_n$ *son los valores de una v.a.* L *de truncamiento por la izquierda.*
- $\delta_1, \delta_2, \ldots, \delta_n$ *son las observaciones de una v.a.* $\delta = I[Z = T]$ *donde* $Z = T \wedge C$. *A esta variable se le suele llamar indicador de censura.*
- $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ *son observaciones de un vector aleatorio de covariables* $\boldsymbol{X}$.
- **Censura independiente.** *Para un vector de covariables específico* $\boldsymbol{x}$, *se supone que* T *y* C *son independientes dado* $\boldsymbol{X} = \boldsymbol{x}$.

Bajo la hipótesis de censura aleatoria independiente, condicionado a las covariables, los elementos censurados son representativos de aquellos que aún están en riesgo en el mismo instante. En otras palabras, las tasas de fallo de los individuos que no están en riesgo, son las mismas que si no estuvieran censuradas, y por lo

tanto, condicionado a las covariables, los elementos no están siendo censurados porque tienen un mayor o menor riesgo de fallo. Para más detalle, ver Kalbfleisch y Prentice (2002).

También se supone de manera implícita que la censura es *no-informativa* en el modelo. Esto es, la función de distribución de la variable de censura no contiene ninguna información sobre la distribución desconocida de los tiempos de vida. Bajo una aproximación paramétrica, esto quiere decir que la distribución de la censura no involucra a ninguno de los parámetros desconocidos en el modelo.

Posiblemente, el modelo semi-paramétrico más conocido para la función de azar condicionada es el extensamente utilizado *modelo de riesgos proporcionales de Cox*, que supone la proporcionalidad de las funciones de azar de dos elementos definidos mediante diferentes conjuntos de covariables. Esta hipótesis puede ser bastante restrictiva en muchos casos, especialmente para la modelización a largo plazo. Por tanto, se han propuesto recientemente varias alternativas al modelo de Cox, en las cuales, no se exige que se cumpla la hipótesis de riesgos proporcionales.

Las alternativas más populares en el campo de la Fiabilidad y la Supervivencia son *el modelo aditivo de Aalen* y *el modelo AFT*. Por otro lado, la situación menos restrictiva surge cuando no se considera ninguna estructura en la función (2.1). Dado un vector de covariables, la estimación no paramétrica de la tasa de riesgo puede abordarse de varias maneras. La metodología más común es utilizando técnicas de suavizado al estimador de Nelson-Aalen (dado un vector de covariables), si bien, otras aproximaciones han desarrollado estimadores de la tasa de riesgo condicionada mediante el cociente de estimadores no paramétricos de la función de densidad condicionada y de la función de supervivencia. A continuación se realiza un breve repaso de estos modelos.

2.2.1. Modelo de riesgos proporcionales (PH)

El modelo de PH de Cox (1972) es el modelo utilizado en la mayoría de las aplicaciones en Bioestadística y, generalmente, en los estudios de Fiabilidad y Supervivencia. En el contexto de la Fiabilidad, el tiempo de supervivencia se interpreta como el tiempo antes de que suceda el fallo en un dispositivo dado (sistema o componente), y el objetivo principal es evaluar este tiempo en términos de las características particulares del dispositivo.

Sea T una v.a. que indica el tiempo de fallo y sea a su vez $\mathbf{X} = (X_1, X_2, \ldots, X_p)^\top$ un vector p-dimensional de covariables o de variables explicativas que describe a un individuo en particular o un sistema en términos de sus características particulares. El modelo de riesgos proporcionales básico supone que la función de intensidad del tiempo de vida de un sujeto con vector de covariables dado por $\mathbf{X}$ puede expresarse como

$$\lambda(t) = Y(t)\lambda_0(t)\Psi(\boldsymbol{\beta}\mathbf{X}), \tag{2.2}$$

donde nuevamente, $Y(t)$ es el *proceso de riesgo*, $\lambda_0(t)$ es una función que depende del tiempo, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ es un vector de parámetros p-dimensional y $\Psi(\cdot)$ es una función desconocida. El modelo no supone ninguna forma paramétrica en particular para la *función de azar base* $\lambda_0(t)$. Esta función expresa la tasa de riesgo de un individuo con nivel de covariables considerado *base*, esto es, representa el riesgo de un elemento con vector de covariables al nivel de referencia (generalmente se considera este nivel como el 0, siempre que $\Psi(0) = 1$). Dado que no se especifica una forma funcional para la función $\lambda_0(t)$, en este modelo no se hacen hipótesis sobre la distribución del tiempo de vida de la población de los elementos base (o de referencia). Por otro lado, se trata de un modelo semi-paramétrico ya que se supone una forma paramétrica para el efecto de las covariables. De hecho, una elección habitual consiste en que $\Psi(\boldsymbol{\beta}\mathbf{X})$ sea

$$\Psi(\boldsymbol{\beta}\mathbf{X}) = \exp(\boldsymbol{\beta}\mathbf{X}) = \exp\left(\sum_{j=1}^{p} \beta_j X_j\right). \tag{2.3}$$

Básicamente, el modelo supone que existe una relación de proporcionalidad entre las funciones de azar de los tiempos de supervivencia correspondiente a diferentes elementos de la población. En otras palabras, si se consideran dos sujetos definidos respectivamente mediante los vectores de covariables $\mathbf{X}^1$ y $\mathbf{X}^2$, la razón de las funciones de azar correspondientes es

$$\frac{\lambda(t;\mathbf{X}^1)}{\lambda(t;\mathbf{X}^2)} = \frac{Y(t)\lambda_0(t)\exp\left(\sum_{j=1}^{p}\beta_j X_j^1\right)}{Y(t)\lambda_0(t)\exp\left(\sum_{j=1}^{p}\beta_j X_j^2\right)} = \exp\left[\sum_{j=1}^{p}\beta_j\left(X_j^1 - X_j^2\right)\right] \tag{2.4}$$

que es constante en el tiempo. La razón de riesgos (2.4), es referida en el contexto de la Bioestadística, como el *riesgo relativo* de un individuo con factor de riesgo $\mathbf{X}^1$ con respecto al evento de interés (por ejemplo, muerte o recaída) cuando es comparado con un individuo con factor de riesgo $\mathbf{X}^2$.

El interés principal es hacer inferencia sobre el vector de parámetros $\boldsymbol{\beta}$ y la función de azar base $\lambda_0(t)$.

En el contexto de este capítulo, se supone que se tienen n observaciones independientes $(L_i, Z_i, \delta_i, \mathbf{X}_i)$, siendo $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^\top$ con $i = 1, 2, \ldots, n$, bajo las especificaciones de un modelo con censura por la derecha y truncamiento por la izquierda. Para cada sujeto $i = 1, 2, \ldots, n$, L_i es el tiempo de truncamiento a la izquierda, $Z_i = T_i \wedge C_i$ es el tiempo de seguimiento (que se suponen ordenados en orden creciente), T_i es el tiempo de vida y C_i es el correspondiente tiempo de censura; el indicador de censura es δ_i, que indica si una observación está censurada o no lo está ($\delta_i = 1$ si ha sucedido el fallo en Z_i y $\delta_i = 0$ si el tiempo de vida está censurado por la derecha) y $\mathbf{X}_i$ es un vector de variables explicativas.

La estimación de los parámetros $\boldsymbol{\beta}$ se ha basado tradicionalmente en una formulación de *verosimilitud parcial o condicionada*, donde la función de azar base se entiende como un "parámetro incómodo", que en muchos casos no se estima ya que el objetivo es evaluar el efecto que cada factor o covariable tiene sobre el riesgo de fallo.

A continuación, se considera un enfoque alternativo de estimación de los pará-

metros basado en el método de mínimos cuadrados.

Siguiendo las directrices del capítulo anterior, los tiempos de supervivencia se formulan usando procesos de recuento. De esta manera, las observaciones pueden representarse como una terna $(N_i(t), Y_i(t), \mathbf{X}_i)$ con $i = 1, 2, \ldots, n$ y $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$ donde $N_i(t)$ es un proceso de recuento con función intensidad especificada de acuerdo con el modelo de Cox (2.2) e $Y_i(t)$ es el proceso de riesgo del sujeto definido como $Y_i(t) = I[L_i \leq t \leq Z_i]$.

Los parámetros de regresión $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ pueden estimarse, por ejemplo, por el principio de mínimos cuadrados que a continuación se explica (para más detalles ver Martinussen y Scheike (2006)).

Sean los vectores

$$\mathbf{N}(t) = (N_1(t), N_2(t), \ldots, N_n(t))^\top, \qquad \mathbf{Y}(t) = (Y_1(t), Y_2(t), \ldots, Y_n(t))^\top$$

un proceso de recuento multivariante y el proceso de riesgo asociado respectivamente. La función intensidad puede expresarse como

$$\boldsymbol{\lambda}(t) = (\lambda_1(t), \lambda_2(t), \ldots, \lambda_n(t))^\top,$$

y del mismo modo, las covariables pueden organizarse en una matriz de diseño de dimensión $p \times n$, de la forma

$$\underline{\mathbf{X}}(t) = \begin{pmatrix} Y_1(t)X_{11} & Y_2(t)X_{21} & \ldots & Y_n(t)X_{n1} \\ Y_1(t)X_{12} & Y_2(t)X_{22} & \ldots & Y_n(t)X_{n2} \\ \vdots & \vdots & & \vdots \\ Y_1(t)X_{1p} & Y_2(t)X_{2p} & \ldots & Y_n(t)X_{np} \end{pmatrix}.$$

Denotando a las intensidades acumuladas n-dimensionales como $\boldsymbol{\Lambda}(t) = \int_0^t \boldsymbol{\lambda}(s)ds$, se tiene que $\mathbf{M}(t) = \mathbf{N}(t) - \boldsymbol{\Lambda}(t)$ es una martingala local de cuadrado integrable n-dimensional. Usando el teorema de descomposición de Doob-Meyer se tiene

$$d\mathbf{N}(t) = \boldsymbol{\lambda}(t)dt + d\mathbf{M}(t) = \operatorname{diag}\left(\exp\left(\boldsymbol{\beta}\underline{\mathbf{X}}_i(t)\right)\right)\mathbf{Y}(t)d\boldsymbol{\Lambda}_0(t) + d\mathbf{M}(t). \qquad (2.5)$$

Dado que los incrementos de las martingalas son incorrelados y tienen media 0, la expresión (2.5) sugiere estimar los parámetros del modelo a partir de ecuaciones de mínimos cuadrados; ver Martinussen y Scheike (2006), donde se procede en primer lugar a resolver el problema para un $\boldsymbol{\beta}$ fijo, obteniendo como estimador de la función de azar acumulada base

$$\widehat{\boldsymbol{\Lambda}}_0(t) = \int_0^t \mathbf{Y}^-(s) d\mathbf{N}(s), \tag{2.6}$$

donde $\mathbf{Y}^-(t) = \left(\mathbf{Y}^\top(t)\text{diag}\left(\exp\left(\boldsymbol{\beta}\underline{\mathbf{X}}_i(t)\right)\right)\mathbf{Y}(t)\right)^{-1}\mathbf{Y}^\top(t)$ es la inversa generalizada de $\mathbf{Y}(t)$, que se supone 0 cuando dicha inversa no existe. Este estimador se sustituye en la correspondiente ecuación *score* para resolver en $\boldsymbol{\beta}$, de modo que el estimador del vector de coeficientes se obtiene resolviendo la ecuación

$$\int \underline{\mathbf{X}}(t)\left(d\mathbf{N}(t) - \text{diag}\left(\exp\left(\boldsymbol{\beta}\underline{\mathbf{X}}_i(t)\right)\right)\mathbf{Y}(t)\mathbf{Y}^-(t)d\mathbf{N}(t)\right) = 0.$$

Bajo determinadas condiciones (ver Martinussen y Scheike, 2006), puede comprobarse que $\widehat{\boldsymbol{\beta}}$ es un estimador consistente.

A partir de $\widehat{\boldsymbol{\beta}}$, se obtiene el estimador de Breslow (1972) para la función de azar acumulado base, es decir

$$\widehat{\Lambda}_0(t) = \widehat{\Lambda}_0(t, \widehat{\boldsymbol{\beta}}) = \int_0^t \frac{dN^{(n)}(s)}{S^0(s, \widehat{\boldsymbol{\beta}})},$$

siendo $S^0(t, \boldsymbol{\beta}) = \sum_{i=1}^n Y_i(t)\exp(\boldsymbol{\beta}\mathbf{X}_i(t))$.

Diagnósticos de regresión

Una vez estimado el modelo cabe preguntarse por la idoneidad del mismo. En este sentido, una cuestión fundamental es evaluar si la forma funcional con la que cada covariable es introducida en el modelo (2.2) está o no correctamente especificada, o puede que los coeficientes de regresión no sean constantes a través del tiempo. Es decir en muchos casos la llamada *hipótesis de PH* puede ser cuestionable.

- Uno de los procedimientos más simples para verificar la hipótesis de PH, es explorar gráficamente la función de azar base acumulado estimado en el modelo de riesgos proporcionales estratificado, basándose en una covariable X_1

$$\lambda(t) = Y(t)\lambda_{0k}(t)\exp\left(\beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p\right),$$

cuando $X_1 = k$ con $k = 1, 2, \ldots, D$, siendo D el posible número de estratos. Si el modelo de Cox es correcto, entonces se estiman los riesgos base acumulados $\widehat{\Lambda}_{0k}(t)$ del modelo de Cox en los distintos estratos, y dichos riesgos base acumulados deberían ser aproximadamente proporcionales. Normalmente se hacen las gráficas de $\log(\widehat{\Lambda}_{0k}(t))$ frente a t, con $k = 1, 2, \ldots, D$ y se comprueba si dichas curvas son aproximadamente paralelas.

- Otro procedimiento para evaluar al modelo de Cox se basa en los residuos martingala definidos como

$$\widehat{M_i}(t) = N_i(t) - \int_0^t Y_i(s)\exp(\widehat{\boldsymbol{\beta}}\mathbf{X}_i)d\widehat{\Lambda}_0(s),$$

donde $\widehat{M_i}(\cdot)$ es una martingala, $\widehat{\boldsymbol{\beta}}$ un estimador de $\boldsymbol{\beta}$ y Λ_0 el riesgo acumulado calculado por el estimador del riesgo base de Breslow.

Este residuo puede ser interpretado, para cada t, como la diferencia en el intervalo $[0, t]$ del número de eventos observados menos los esperados proporcionados por el modelo, es decir, el exceso de muertes (fallos) observados. Los residuos poseen algunas de las propiedades de los modelos lineales, como son $\sum_{i=1}^{n} \widehat{M_i}(t) = 0$ para cualquier t, y $E[\widehat{M_i}(t)] = \text{Cov}\left[\widehat{M_i}(t), \widehat{M_j}(t)\right] = 0$ asintóticamente.

Para el modelo de PH con covariables que no dependen del tiempo, donde t_i denota el tiempo de observación para el sujeto i y δ_i el estado final, este residuo puede reducirse a una expresión más simple de la forma

$$\widehat{M_i} = \delta_i - \exp(\widehat{\boldsymbol{\beta}}\mathbf{X}_i)\widehat{\Lambda}_0(t_i).$$

Los residuos de martingala suelen ser de carácter fuertemente asimétricos y poseen una cola muy larga hacia la derecha.

- Otra forma para explorar la hipótesis del modelo PH es a través de los residuos de Schoenfeld (1982), que se definen como la matriz

$$r_{ij}(\boldsymbol{\beta}) = X_{ij} - \left(\frac{S^1(t_i, \boldsymbol{\beta})}{S^0(t_i, \boldsymbol{\beta})}\right)^{\otimes 2}$$

que tiene una fila por fallo (muerte) y una columna por covariable, donde i y t_i son los sujetos y el tiempo de ocurrencia del evento respectivamente, $S^0(t, \boldsymbol{\beta})$ y $S^1(t, \boldsymbol{\beta})$ se definen como $S^0(t, \boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i(t) \exp(\boldsymbol{\beta}\mathbf{X}_i)$ y $S^1(t, \boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i(t) \exp(\boldsymbol{\beta}\mathbf{X}_i)\mathbf{X}_i^\top$.

Bajo el supuesto de riesgos proporcionales, los residuos de Schoenfeld siguen un patrón aleatorio, por lo tanto, son útiles en la evaluación de la tendencia a través del tiempo o, dicho de otro modo, de la falta de proporcionalidad de riesgos. Therneau y Lumley (2008), consideran que los coeficientes de regresión vienen dados mediante funciones dependientes del tiempo de la forma $\beta(t) = \beta + \theta g(t)$, para una función suave $g(t)$. Si la regresión lineal generalizada de los residuos de Schoenfeld en función del tiempo proporciona una pendiente distinta de cero podría estar indicando la violación del supuesto de riesgos proporcionales, ya que la hipótesis nula de riesgos proporcionales sería equivalente a $\theta = 0$.

Al igual que cualquier tipo de regresión, se recomienda examinar la gráfica de la recta de regresión obtenida, además de realizar test para comprobar si la pendiente sea distinta de cero.

2.2.2. Modelo de riesgos aditivos de Aalen

Cuando el modelo de riesgos proporcionales de Cox no se ajusta apropiadamente a los datos, una alternativa viene dada por el modelo aditivo propuesto por Aalen

(1978). El modelo establece que la intensidad de un proceso de recuento $N(t)$, $t \in [0, \tau]$, $\tau < \infty$ de un sujeto con covariables $\mathbf{X}(t)$ (acotadas y predecibles) es de la forma

$$\boldsymbol{\lambda}(t) = \mathbf{Y}(t)\boldsymbol{\beta}(t)\mathbf{X}(t) \tag{2.7}$$

donde $\mathbf{Y}$(t) es el indicador de riesgo, $\mathbf{X}(t) = (X_1(t), X_2(t), \ldots, X_p(t))^\top$ indica un vector de posibles covariables dependientes del tiempo y $\boldsymbol{\beta}(t) = (\beta_1(t), \beta_2(t), \ldots, \beta_p(t))$ son los coeficientes de regresión $\left(\int_0^t |\beta_j(s)|\, ds < \infty,\ j = 1, 2, \ldots, p\right)$.

El problema de estimar los coeficientes de regresión acumulados, definidos como $B(t) = \int_0^t \beta(s)ds$, resulta más fácilmente abordable.

Sea $(N_i(t), Y_i(t), \mathbf{X}_i)$ con $i = 1, 2, \ldots, n$ y $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$, una muestra de n sujetos tales que la intensidad $\lambda_i(t)$ para el proceso de recuento i-ésimo $N_i(t)$ es de la forma (2.7). Se define a su vez

$$\mathbf{N}(t) = (N_1(t), N_2(t), \ldots, N_n(t))^\top, \qquad \boldsymbol{\lambda}(t) = (\lambda_1(t), \lambda_2(t), \ldots, \lambda_n(t))^\top,$$

el proceso de recuento n-dimensional considerando todos los sujetos y sus intensidades, respectivamente. Si se supone un contexto de datos filtrados (con censura por la derecha y truncamiento por la izquierda), el proceso de riesgo viene definido de la forma $Y_i(t) = I[L_i \leq t \leq Z_i]$, $i = 1, 2, \ldots, n$, se define la matriz $\underline{\mathbf{X}}(t)$ de dimensiones $p \times n$ de la misma forma que en el modelo de Cox del apartado anterior.

Sea $\boldsymbol{\Lambda}(t) = \int_0^t \boldsymbol{\lambda}(s)ds$ las intensidades acumuladas n-dimensionales tal que $\mathbf{M}(t) = \mathbf{N}(t) - \boldsymbol{\Lambda}(t)$ es una martingala n-dimensional, se tiene que

$$d\mathbf{N}(t) = \boldsymbol{\lambda}(t)dt + d\mathbf{M}(t) = \boldsymbol{\beta}(t)\underline{\mathbf{X}}(t)dt + d\mathbf{M}(t), \tag{2.8}$$

y puesto que los incrementos de las martingalas son incorrelados y tienen de media cero, esta ecuación sugiere que los incrementos de $\boldsymbol{\beta}(t)dt$, que pueden escribirse como $d\mathbf{B}(t)$, pueden estimarse mediante técnicas de regresión lineal múltiple. Para resolver el problema de regresión lineal múltiple se define la inversa generalizada

de $\underline{\mathbf{X}}(t)$ como la matriz de dimensiones $p \times n$ dada por

$$\underline{\mathbf{X}}^{-}(t) = \left(\underline{\mathbf{X}}(t)\mathbf{W}(t)\underline{\mathbf{X}}^{\top}(t)\right)^{-1}\underline{\mathbf{X}}(t)\mathbf{W}(t),$$

donde $\mathbf{W}(t)$ es una matriz de pesos diagonal y predecible de dimensiones $n \times n$. Imponiendo que $\underline{\mathbf{X}}^{-}(t)$ es igual a cero cuando la inversa no exista, y sea $J(t)$ igual a 1 cuando la inversa exista y cero en otro caso, la inversa generalizada satisface la relación

$$\underline{\mathbf{X}}^{-}(t)\underline{\mathbf{X}}^{\top}(t) = J(t)\mathbf{I}_p,$$

donde $\mathbf{I}_p$ es la matriz identidad de dimensión p. La expresión (2.8) conduce al estimador que puede escribirse en forma de integral como

$$\widehat{\mathbf{B}}(t) = \int_0^t \underline{\mathbf{X}}^{-}(s)d\mathbf{N}(s). \tag{2.9}$$

Bajo apropiadas condiciones de regularidad (ver Martinussen y Scheike, 2006), puede comprobarse la consistencia y normalidad asintótica de este estimador. A partir de aquí un estimador $\widehat{\boldsymbol{\beta}}(t)$ puede obtenerse utilizando técnicas de suavizado con funciones núcleo, es decir, $\widehat{\boldsymbol{\beta}}(t) = \displaystyle\int_0^\tau \frac{1}{h}K\left(\frac{t-u}{h}\right)d\widehat{\mathbf{B}}(u)$ con K una función núcleo y h un parámetro de suavizado.

2.2.3. Modelo de tiempo de vida acelerada (AFT)

El modelo AFT que va a tratarse, es un modelo de regresión que relaciona el logaritmo del tiempo de vida T de manera lineal con el vector de covariables $\mathbf{X} = (X_1, X_2, \ldots, X_p)^{\top}$. Específicamente puede escribirse de la forma

$$\log T = -\boldsymbol{\beta}\mathbf{X} + \varepsilon,$$

donde ε es un término de error aleatorio y $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ es un vector con los parámetros de regresión. Esta especificación conduce a la siguiente función de azar para T dado $\mathbf{X}$

$$\lambda(t) = \lambda_0\left(t\exp(\boldsymbol{\beta}\mathbf{X})\right)\exp(\boldsymbol{\beta}\mathbf{X}), \tag{2.10}$$

donde $\lambda_0(t)$ es la función de azar asociada a la variable $T_0 = \exp(\varepsilon)$.

Como puede verse las covariables actúan de manera multiplicativa en el tiempo por lo que sus efectos son los de acelerar o decelerar el tiempo de fallo en relación a T_0.

El ajuste de este modelo resulta algo más complicado con observaciones censuradas y las propiedades asintóticas de los estimadores son también más difíciles de obtener. No obstante han sido muchos los trabajos desarrollados en este tema durante los últimos años dotando al modelo de una mayor aplicabilidad práctica. En la Sección 2.3 se tratará este tema con más detalle.

2.2.4. Regresión no paramétrica de la función de azar

La estimación no paramétrica de la función de azar usando suavizados de tipo núcleo puede llevarse a cabo a partir de dos enfoques diferentes, que dan lugar a dos familias de estimadores no paramétricos para la función de azar, a saber: estimadores "externos" y estimadores "internos", tal y como vienen clasificados en Nielsen y Linton (1995) siguiendo la notación introducida por Jones, Davies y Park (1994).

- "Estimadores externos". Son estimadores dados por la definición de la función de azar como el cociente entre la función de densidad y la función de supervivencia $\alpha(t) = f(t)/S(t)$. Esta familia de estimadores resultan de sustituir estimaciones no paramétricas de la función de densidad y la función de supervivencia $\widehat{f}$ y $\widehat{S}$, respectivamente.

- "Estimadores internos". Surgen del suavizado de los incrementos de la estimación no paramétrica de la función de azar acumulada $\widehat{\Lambda}$. Por ejemplo, se considera el siguiente estimador tipo núcleo de la función de azar

$$\widehat{\alpha}(t) = \int K_h(t-u)d\widehat{\Lambda}(u),$$

donde $h > 0$ es el parámetro de suavizado y $K_h = (1/h)K(\cdot/h)$, siendo K una función tipo núcleo.

Los métodos mencionados en este apartado se centran en el primer enfoque, dando así un estimador local para la densidad y la función de supervivencia condicionadas. Estos métodos han sido sugeridos por Spierdijk (2008) y pueden considerarse una simplificación del estimador lineal local introducido por Nielsen (1998a) en el contexto de la regresión de la función de azar.

Análogamente a secciones anteriores, se considera una muestra de observaciones i.i.d. de tamaño n de la forma $\{(L_i, Z_i, \delta_i, \mathbf{X}_i)\,,\, i = 1, 2, \ldots, n\}$, donde L_i es el tiempo de truncamiento y $Z_i = T_i \wedge C_i$ es el tiempo de observación. Para un perfil de covariables dado por el vector $\mathbf{X}$, sea $H(t;\mathbf{X}) = (1 - F(t;\mathbf{X}))(1 - G(t;\mathbf{X}))$ la función de supervivencia correspondiente a Z, donde F y G son las funciones de distribución de T y C respectivamente. Por tanto, si $G(t;\mathbf{X}) < 1$, la función de azar se puede expresar como

$$\alpha(t;\mathbf{X}) = \frac{f(t;\mathbf{X})}{1 - F(t;\mathbf{X})}\frac{1 - G(t;\mathbf{X})}{1 - G(t;\mathbf{X})} = \frac{f(t;\mathbf{X})(1 - G(t;\mathbf{X}))}{H(t;\mathbf{X})}. \tag{2.11}$$

Si se considera la distribución conjunta del vector aleatorio (Z, δ), para $\delta = 1$

$$g(t, \delta = 1) = \lim_{\Delta \to 0} Pr\{t < Z \leq t + \Delta \mid \delta = 1\} = \lim_{\Delta \to 0} Pr\{t \leq T \leq t + \Delta \mid C > t + \Delta\}.$$

Puesto que T y C se suponen independientes, la expresión anterior da lugar a

$$g(t, \delta = 1) = \lim_{\Delta \to 0} Pr\{t < T \leq t + \Delta\} \lim_{\Delta \to 0} Pr\{C > t + \Delta\} = f(t)(1 - G(t)),$$

y entonces, la expresión (2.11) se reduce a

$$\alpha(t;\mathbf{X}) = \frac{g(t, \delta = 1;\mathbf{X})}{H(t;\mathbf{X})}.$$

Por tanto, para obtener un estimador no paramétrico de la función de azar condicionada, se sustituyen $g(t, \delta = 1;\mathbf{X})$ y $H(t;\mathbf{X})$ por las correspondientes estimaciones no paramétricas. Para ello, se construyen estimaciones suaves para cada uno de los dos términos que se denotan $\widehat{g}(t, \delta = 1;\mathbf{X})$ y $\widehat{H}(t;\mathbf{X})$.

Para estimar g, sólo se tienen en cuenta las observaciones no censuradas, mientras que para el estimador de la función del denominador H, se hace uso de todas las observaciones, estén censuradas o no. Condicionado al evento $\{\delta = 1\}$, puede procederse de la siguiente forma

$$g(t, \delta = 1) = \lim_{\Delta \to 0} Pr\{t < Z \leq t + \Delta \mid \delta = 1\} Pr\{\delta = 1\} = g_1(t) Pr\{\delta = 1\}$$

de modo que basta definir los estimadores de los dos factores de esta última expresión. Un estimador natural de $Pr\{\delta = 1\}$ viene dado por n_1/n, donde n_1 es el número de observaciones no censuradas en la muestra. Para estimar g_1, basta considerar la submuestra no censurada, a partir de la cual se construye el estimador núcleo de la densidad

$$\widehat{g}_1(t) = \frac{1}{n_1} \sum_{i \in U} K_{h_1}(t - T_i), \tag{2.12}$$

donde $K_{h_1}(\cdot) = (1/h_1)K(\cdot/h_1)$ y $U \subseteq \{1, 2, \ldots, \}$ el conjunto de índices que corresponden a las observaciones no censuradas.

El estimador (2.12) aplica un peso de $1/n_1$ a cada observación muestral no censurada. Este es el estimador de tipo núcleo de la densidad g_1 en el caso en el que no existan covariables. Considerando la información dada por las variables explicativas, estos pesos pueden sustituirse por los pesos de Nadaraya-Watson, Nadaraya (1964) y Watson (1964), obteniendo

$$\widehat{g}_1(t; x) = \sum_{i \in U} K^*_{h_2,i}(x) K_{h_1}(t - T_i), \tag{2.13}$$

donde

$$K^*_{h_2,i}(x) = \frac{K_{h_2}(x - X_i)}{\sum_{j \in U} K_{h_2}(x - X_j)} = \frac{K\left(\dfrac{x - X_i}{h_2}\right)}{\displaystyle\sum_{j \in U} K\left(\frac{x - X_j}{h_2}\right)}, \ \forall i \in U. \tag{2.14}$$

El estimador resultante de la función de azar no condicionada está íntimamente relacionado con el sugerido en Nielsen y Linton (1995), pero ahora no se supone

que la función núcleo para el suavizado en la variable de tiempo K_1 sea necesariamente la misma que para las covariables K_2, además, tampoco se supone que los parámetros de suavizado h_1 y h_2 sean diferentes.

Los pesos podrían también haber sido elegidos basándose en un procedimiento de suavizado lineal local como el que se desarrolla en Spierdijk (2008).

Por último, el estimador correspondiente para la función de supervivencia H puede obtenerse de forma similar mediante

$$\widehat{H}(t;x) = \sum_{i=1}^{n} K^*_{h_2,i}(x) K_{h_1}(t - T_i),$$

donde hay que resaltar que todas las observaciones $\{(Z_i, X_i);\, i = 1, 2, \ldots, n\}$ son ahora utilizadas en el procedimiento de estimación, independientemente de si están censuradas o no.

El estimador lineal local considerado antes supone una simplificación del propuesto en Nielsen (1998b), donde se lleva a cabo un ajuste lineal local multivariante basado en mínimos cuadrados para el problema de regresión de la función de azar. Además se obtienen resultados acerca de la convergencia puntual de los estimadores (constante y lineal) locales. En Spierdijk (2008), la aproximación lineal local se desarrolla sólo para el suavizado que involucra a los argumentos de la covariable, mientras que para la variable de tiempo, se utiliza el estimador constante local correspondiente de la densidad de tipo núcleo en el numerador y el estimador de tipo núcleo de la función de supervivencia en el denominador.

El problema de estimación suave no está resuelto hasta que un valor del parámetro de suavizado apropiado es seleccionado. Existen varios métodos para la selección de dicho parámetro con buenas propiedades asintóticas en el contexto de datos completos, sin embargo en los modelos de regresión con datos de tiempos de vida, los métodos de selección del parámetro de suavizado están aún en una etapa inicial. Spierdijk (2008) propone un procedimiento de selección del parámetro de suavizado basado en métodos plug-in. Sin embargo, la selección local y global del

parámetro de suavizado óptimo para la regresión de la función de azar sigue siendo, a día de hoy, un problema de estudio (ver Gámiz-Pérez, Janys, Martínez-Miranda y Nielsen(2013b)).

2.3. Estimación no paramétrica del modelo AFT

Igual que en las secciones anteriores, sea $T_i,\ i = 1, 2, \ldots, n$, el tiempo de fallo para un sujeto i-ésimo, suponiendo que los sujetos son independientes, $N_i(t)$ es el número de fallos que suceden en el sujeto i-ésimo en el instante t e $Y_i(t)$ el proceso de riesgo. Se define la función media del proceso de recuento $N_i(t)$ condicionado al vector de covariables $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^\top$ como

$$E\left[N_i(t)|\mathbf{X}_i\right] = Pr\left\{T_i \leq t|\mathbf{X}_i\right\} = \Psi_0\left(t\exp\left(-\boldsymbol{\beta}\mathbf{X}_i\right)\right), \tag{2.15}$$

donde $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ es un vector de parámetros desconocidos, y Ψ_0 es una función continua y no especificada. Si se escribe $T_{0,i} = T_i \exp\left(-\boldsymbol{\beta}\mathbf{X}_i\right)$ y se define $N_{0,i}(t) = I\left[T_{0,i} \leq t\right]$, entonces, claramente $N_{0,i}(t) = N_i\left(t\exp\left(\boldsymbol{\beta}\mathbf{X}_i\right)\right)$, y también, utilizando la expresión (2.15) se obtiene que

$$E\left[N_{0,i}(t)\right] = Pr\left\{T_{0,i} \leq t\right\} = \Psi_0(t).$$

Esto quiere decir que la probabilidad de fallo en el tiempo t cuando $\mathbf{X}_i = \mathbf{x}$ es igual a la probabilidad de fallo en el tiempo $(t\exp(-\boldsymbol{\beta}\mathbf{x}))$ cuando $\mathbf{X}_i = 0$. En otras palabras, el conjunto de covariables $\mathbf{X}_i$ afecta a la probabilidad de ocurrencia de fallo expandiendo o comprimiendo la escala de tiempo, en la cual este suceso ocurre por un factor multiplicativo en relación al tiempo que le correspondería a un sujeto con valor de covariables 0. Dicho factor multiplicativo es del tipo $\exp(-\boldsymbol{\beta}\mathbf{x})$, por lo tanto, se cumple la siguiente relación

$$T_i = T_{0,i}\exp(\boldsymbol{\beta}\mathbf{X}_i), \tag{2.16}$$

o, equivalentemente, puede formularse el siguiente modelo de regresión log-lineal

$$\log T_i = \boldsymbol{\beta}\mathbf{X}_i + \varepsilon_i, \tag{2.17}$$

donde el término de error $\varepsilon_i = \log T_{0,i}$, para todo $i = 1, 2, \ldots, n$, tiene una distribución con función de supervivencia S_0. A esta función se le denomina función de supervivencia base, y representa la función de supervivencia de un sujeto cuando las covariables tienen valor en el nivel base (cero).

Como se definió antes, el esquema de censura por la derecha y/o truncamiento por la izquierda puede formularse considerando un caso particular de los procesos de recuento.

Sea F la función de distribución del tiempo de fallo T y se asume el modelo (2.17). Si ocurre la censura por la derecha y el truncamiento por la izquierda, entonces, puede considerarse el modelo LTRC dado en la Definición 20 de la Sección 2.2.

Se recuerda que el modelo LTRC puede formularse usando un caso particular de los procesos de fallo N y de riesgo Y definidos anteriormente. En concreto, para cada individuo $i = 1, 2, \ldots, n$, el correspondiente proceso de fallo viene dado por $N_i(t) = I\left[Z_i \leq t\right] \delta_i$, y el proceso de riesgo $Y_i(t) = I\left[L_i \leq t \leq Z_i\right]$, para $t \geq 0$. La formulación del proceso de recuento supuesta en este trabajo es, de hecho, una formulación más general que incorpora el caso LTRC, como un caso particular.

A partir de la ecuación (2.16) o (2.17) se puede evaluar el efecto que las covariables tiene sobre el tiempo de vida de un sujeto en particular. Sin embargo, el interés de este capítulo va más allá de evaluar los efectos y se trata de calcular la probabilidad de que un sujeto en particular sobreviva más de un tiempo específico t. En otras palabras, se desea estimar la siguiente función

$$\begin{aligned} S_x(t) &= Pr\left\{T_x > t\right\} = Pr\left\{T_0 \exp\left(\boldsymbol{\beta}\mathbf{x}\right) > t\right\} = Pr\left\{T_0 > t \exp\left(-\boldsymbol{\beta}\mathbf{x}\right)\right\} \\ &= S_0\left(t \exp\left(-\boldsymbol{\beta}\mathbf{x}\right)\right), \end{aligned} \tag{2.18}$$

donde S_0 es la función de supervivencia base. Cabe destacar que $S_x(\cdot)$ es la función de supervivencia para un sujeto cuando $\mathbf{X}_i = \mathbf{x}$, mientras que $S_0(\cdot)$ representa la función de supervivencia de un individuo cuando las covariables son cero (función base).

Para estimar la función (2.18), se propone un procedimiento secuencial. En primer lugar, se estima el vector de los coeficientes de regresión $\boldsymbol{\beta}$ en el modelo (2.17) utilizando la aproximación semi-paramétrica utilizada por Stute (1996a, 1996b) y Gross y Lai (1996). A partir del estimador $\widehat{\boldsymbol{\beta}}$, se transforman los datos a la escala de tiempos base. Seguidamente, a partir de los datos transformados, se estima la función de supervivencia base mediante un estimador no paramétrico $\widehat{S}_0(\cdot)$. Concretamente, se considera la aproximación lineal local de Nielsen y Tanggaard (2001) y Nielsen *et al.* (2009). Finalmente, la función de supervivencia para un sujeto cuando $\mathbf{X}_i = \mathbf{x}$ se estima de la forma

$$\widehat{S}_x(t) = \widehat{S}_0\left(t \exp(-\widehat{\boldsymbol{\beta}}\mathbf{x})\right), \tag{2.19}$$

donde se hace una transformación de los tiempos de cada sujeto para llevarlos a su escala original. En las siguientes secciones se verá una descripción más detallada de este procedimiento.

2.3.1. Estimación de los coeficientes de regresión

En esta sección se analiza la influencia del conjunto de variables sobre el tiempo de fallo. Para ello, se considera el modelo de regresión especificado en (2.17) siguiendo la aproximación de Stute (1996a, 1996b) y Gross y Lai (1996) que utilizan la formulación de LTRC descrita en la Sección 2.2. Se considera una muestra aleatoria de tamaño n de la forma $\{(L_i, Z_i, \delta_i, \mathbf{X}_i);\ i = 1, 2, \ldots, n\}$, con $L_i \leq Z_i$, $Z_i = T_i \wedge C_i$ y $\delta_i = I\left[Z_i = T_i\right]$, donde $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^\top$ es un vector con dimensión p de covariables que describen al i-ésimo sujeto. Teniendo en cuenta la formulación descrita en la Sección 2.2, y los métodos descritos en el Capítulo 1, la función de azar acumulado se estima por

$$\widehat{\Lambda}(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(s)}{Y^{(n)}(s)} ds,$$

donde $Y^{(n)}(s) = \sum_{i=1}^{n} Y_i(s)$. El estimador de la función de supervivencia viene dado por el estimador de Kaplan-Meier; ver Fleming-Harrington (1991), por tanto

$$\widehat{S}(t) = \prod_{s \leq t} \left(1 - d\widehat{\Lambda}(s)\right).$$

Computacionalmente, resulta más apropiado escribir la ecuación anterior de la forma

$$\widehat{S}(t) = \prod_{T_i \leq t} \left(1 - \frac{\delta_i}{Y^{(n)}(T_i)}\right).$$

Esta expresión proporciona una función escalonada cuyos saltos sólo se producen en los tiempos de fallo T_i y cuyos tamaños vienen dados por

$$W_i = \widehat{S}(T_{i-1}) - \widehat{S}(T_i) = \prod_{j=1}^{i-1} \left(1 - \frac{\delta_j}{Y^{(n)}(T_j)}\right) \frac{\delta_i}{Y^{(n)}(T_i)}. \tag{2.20}$$

Para obtener el estimador del vector de parámetros $\boldsymbol{\beta}$ para datos con censura por la derecha, Stute (1996a, 1996b) propuso una metodología que requiere hipótesis muy generales y cuyo procedimiento de estimación se basa en el criterio de mínimos cuadrados ponderados. Bajo el modelo (2.17), el estimador de $\boldsymbol{\beta}$ se obtiene minimizando

$$\sum_{i=1}^{n} \mathbf{W}_{(i)} \left(\mathbf{Z}_{(i)} - \boldsymbol{\beta}\mathbf{X}_{(i)}\right)^2, \tag{2.21}$$

donde $\mathbf{Z}_{(i)}$ es el i-ésimo valor ordenado de la variable respuesta observada transformada logarítmicamente, $\mathbf{X}_{(i)}$ es la covariable asociada a $\mathbf{Z}_{(i)}$ y $\mathbf{W}_{(i)}$ son los pesos de Kaplan-Meier que se obtienen como sucesivos incrementos del estimador de Kaplan-Meier, ecuación (2.20).

Minimizando la expresión (2.21) se obtiene el estimador de $\boldsymbol{\beta}$ de dimensión $p \times 1$ dado por

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{Z}, \tag{2.22}$$

donde $\mathbf{Z} = \left(\log Z_{(1)}, \log Z_{(2)}, \ldots, \log Z_{(n)}\right)^\top$, $\mathbf{W}$ es una matriz diagonal con los pesos de Kaplan-Meier (ver Kaplan y Meier (1958)) y $\mathbf{X}$ es una matriz con dimensión $n \times p$

con filas $\mathbf{X}_i$, $i = 1, 2, \ldots, n$. En Stute (1996a, 1996b) se estudiaron la consistencia y las propiedades asintóticas de dicho estimador. Además, se propuso un estimador de tipo Jackknife para calcular la varianza asintótica del estimador.

En un marco similar, Gross y Lai (1996) consideraron el problema de regresión cuando el truncamiento también estaba presente en los datos y los pesos se obtuvieron como en la expresión (2.20). Estos autores demostraron que, bajo ciertas condiciones de regularidad, la solución de $\widehat{\boldsymbol{\beta}}$ definida por (2.22) es fuertemente consistente y asintóticamente normal.

2.3.2. Estimación del error estándar

En el trabajo de Gross y Lai (1996) también se sugiere una metodología bootstrap para estimar el error estándar del estimador $\widehat{\boldsymbol{\beta}}$ como sigue:

1. Se generan B muestras aleatorias de tamaño n con reemplazamiento a partir de la muestra original $\{(L_i, Z_i, \delta_i, \mathbf{X}_i);\ i = 1, 2, \ldots, n\}$.

2. Con cada muestra bootstrap ($b = 1, 2, \ldots, B$) se calcula un estimador (bootstrap) de los parámetros de regresión $\boldsymbol{\beta}$ utilizando la expresión (2.22). Este estimador se denota como $\widehat{\boldsymbol{\beta}}^{(b)}$.

3. Se estima el error estándar de $\widehat{\beta}_j$ ($j = 1, 2, \ldots, p$) mediante

$$\mathrm{se}[\widehat{\beta}_j] = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\beta}_j^{(b)} - \overline{\beta}_j\right)^2},$$

   con $\overline{\beta}_j = \dfrac{1}{B}\displaystyle\sum_{b=1}^{B}\widehat{\beta}_j^{(b)}$.

4. Se calcula el intervalo de confianza al $(1-\alpha)\times 100\,\%$ para β_j ($j = 1, 2, \ldots, p$) de la forma

$$\left(\widehat{\beta}_j \pm z_{\alpha/2} \times \mathrm{se}[\widehat{\beta}_j]\right),$$

   donde $z_{\alpha/2}$ es el cuantil $(1 - \alpha/2)$ de la distribución Normal estándar.

2.3.3. Estimación no paramétrica de la función de supervivencia base

En las secciones previas se han descrito los estimadores para los coeficientes del modelo de regresión AFT. Ahora, para obtener la función de supervivencia (2.19) de un sujeto con $\mathbf{X}_i = \mathbf{x}$, falta estimar la función de supervivencia base $S_0(\cdot)$. Seguidamente, se introduce un estimador lineal local de dicha función. Sin pérdida de generalidad, se supone que las componentes numéricas del vector de covariables tienen media cero, $E[\mathbf{X}] = 0$.

Transformación de los datos

Por simplicidad, se considera la formulación LTRC. Utilizando el estimador $\widehat{\boldsymbol{\beta}}$ del vector de coeficientes $\boldsymbol{\beta}$ del modelo de AFT (2.17), se considera la transformación dada por $u = \exp(-\widehat{\boldsymbol{\beta}}\mathbf{x})y$. Por lo tanto, se construye un nuevo conjunto de datos artificial de la forma

$$\begin{cases} L_{0,i} = \exp(-\widehat{\boldsymbol{\beta}}\mathbf{X}_i)L_i \\ Z_{0,i} = \exp(-\widehat{\boldsymbol{\beta}}\mathbf{X}_i)Z_i \\ \delta_{0,i} = \delta_i \end{cases}$$

donde $L_{0,i} \leq Z_{0,i}$ y $\exp(-\widehat{\boldsymbol{\beta}}\mathbf{X}_i) \geq 0$. Cabe destacar que el conjunto de datos especificado de la forma $\{(L_{0,i}, Z_{0,i}, \delta_{0,i});\ i = 1, 2, \ldots, n\}$ constituye una muestra de tipo LTRC de la población base que puede utilizarse para estimar $S_0(t) = Pr\{T_0 > t\}$.

Considerando la notación utilizada en la Sección 2.2, para cada sujeto $i = 1, 2, \ldots, n$, se define el proceso de fallo como

$$N_{0,i}(t) = \begin{cases} 1; & \text{si } T_{0,i} \leq t \\ 0; & \text{en otro caso} \end{cases},$$

y el proceso de riesgo mediante

$$Y_{0,i}(t) = \begin{cases} 1; & \text{si } L_{0,i} \leq t \leq T_{0,i} \\ 0; & \text{en otro caso} \end{cases}.$$

Con lo cual, el conjunto de datos $\{(L_{0,i}, Z_{0,i}, \delta_{0,i});\ i = 1, 2, \ldots, n\}$ es equivalente a $\{(N_{0,i}(t), Y_{0,i}(t));\ t \geq 0, i = 1, 2, \ldots, n\}$.

En la siguiente sección, se utilizan estos datos transformados para obtener el estimador no paramétrico de la función de supervivencia S_0, utilizando la aproximación lineal local Nielsen y Tanggaard (2001) y Nielsen *et al.* (2009).

Estimador lineal local de la función de supervivencia base

Sea $\widetilde{S}_0(t)$ un estimador empírico de la función de supervivencia S_0, (por ejemplo, el estimador de Nelson-Aalen extendido). Basándose en una aproximación lineal local se obtiene un estimador de $S_0(t)$ como resultado del siguiente problema de mínimos cuadrados ponderados

$$\begin{pmatrix} \widehat{\theta}_0(t) \\ \widehat{\theta}_1(t) \end{pmatrix} = \underset{\theta_0,\theta_1}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \int_0^\tau \left[\widetilde{S}_0(s) - \theta_0 - \theta_1(t-s) \right]^2 K_h(t-s) W(s) Y_i(s) ds \right\},$$

donde $W(\cdot)$ es una función general de pesos. Según la recomendación de Nielsen y Tanggaard (2001), se asume el peso natural $W(s) = 1$. Por tanto, $S_0(t)$ se estima mediante $\widehat{S}_0(t) = \widehat{\theta}_0$.

Resolviendo las siguientes ecuaciones, puede obtenerse una expresión explícita de $\widehat{\theta}_0$.

$$\begin{aligned}
0 &= -2 \sum_{i=1}^{n} \int_0^\tau \left[\widetilde{S}_0(s) - \theta_0 - \theta_1(t-s) \right] K_h(t-s) Y_i(s) ds \\
&= \int_0^\tau \widetilde{S}_0(s) K_h(t-s) Y^{(n)}(s) ds - \theta_0 \int_0^\tau K_h(t-s) Y^{(n)}(s) ds \\
&\quad -\theta_1 \int_0^\tau (t-s) K_h(t-s) Y^{(n)}(s) ds, \\
0 &= -2 \sum_{i=1}^{n} \int_0^\tau \left[\widetilde{S}(s) - \theta_0 - \theta_1(t-s) \right] (t-s) K_h(t-s) Y_i(s) ds \\
&= \int_0^\tau \widetilde{S}(s)(t-s) K_h(t-s) Y^{(n)}(s) ds - \theta_0 \int_0^\tau (t-s) K_h(t-s) Y^{(n)}(s) ds \\
&\quad -\theta_1 \int_0^\tau (t-s)^2 K_h(t-s) Y^{(n)}(s) ds.
\end{aligned}$$

Donde $a_j(t)$ se definió en (1.26) y

$$G_j(t) = \int_0^\tau \widetilde{S}_0(s) K_h(t-s)(t-s)^j Y^{(n)}(s) ds, \quad j = 0, 1.$$

Las ecuaciones anteriores pueden escribirse de forma simplificada como

$$\begin{aligned} G_0(t) &= \theta_0 a_0(t) + \theta_1 a_1(t) \\ G_1(t) &= \theta_0 a_1(t) + \theta_1 a_2(t). \end{aligned}$$

Resolviendo el sistema de ecuaciones se obtiene la siguiente expresión para θ_0

$$\widehat{\theta}_0 = \frac{a_2(t)G_0(t) - a_1(t)G_1(t)}{a_0(t)a_2(t) - a_1^2(t)}.$$

Por tanto, el estimador lineal local de la función de supervivencia puede escribirse explícitamente mediante

$$\widehat{S}_0(t) = \widehat{\theta}_0 = \sum_{i=1}^n \int_0^\tau \left(\frac{a_2(t) - a_1(t)(t-s)}{a_2(t)a_0(t) - a_1^2(t)} \right) K_h(t-s) \widetilde{S}_0(s) Y_i(s) ds.$$

Utilizando la función núcleo lineal local dada por

$$\overline{K}_{t,h}(t-s) = \frac{a_2(t) - a_1(t)(t-s)}{a_2(t)a_0(t) - a_1^2(t)} K_h(t-s), \tag{2.23}$$

El estimador lineal local se puede reescribir de la siguiente forma

$$\widehat{S}_0(t) = \sum_{i=1}^n \int_0^\tau \overline{K}_{t,h}(t-s) \widetilde{S}_0(s) Y_i(s) ds. \tag{2.24}$$

Teniendo en cuenta que la función (2.23) es una función núcleo de segundo orden con respecto a la medida $d\kappa(s) = Y^{(n)}(s)ds$, se tiene que

$$\int_0^\tau \overline{K}_{t,h}(t-s) Y^{(n)}(s) ds = 1, \qquad \int_0^\tau \overline{K}_{t,h}(t-s)(t-s) Y^{(n)}(s) ds = 0,$$
$$\int_0^\tau \overline{K}_{t,h}(t-s)(t-s)^2 Y^{(n)}(s) ds > 0.$$

Propiedades asintóticas del estimador lineal local de la función de supervivencia base

En esta sección se obtienen las propiedades asintóticas del estimador lineal local dado en la expresión (2.24). Para simplificar la notación, se escribe la función de supervivencia base de la forma $S_0(\cdot) = S(\cdot)$ y se denota al estimador lineal local por $\widehat{S}(t)$.

Teorema 4. *Suponiendo las siguientes hipótesis generales:*

(A.1) $h \to 0$ *y* $nh \to \infty$, *cuando* $n \to \infty$.

(A.2) Existe una función positiva γ *tal que el* $\sup_{s\in[0,\tau]} \left\{ \left| Y^{(n)}(s)/n - \gamma(s) \right| \right\} \xrightarrow[n\to\infty]{\mathcal{P}} 0.$

(A.3) $\gamma \in C^{(1)}([0,\tau])$.

(A.4) $\alpha, S \in C^{(2)}([0,\tau])$.

Entonces,

$$B(t) = \frac{1}{2}h^2 S''(t)\mu_2(K) + o_P(h^2),$$

y

$$\sqrt{n}V(t) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \upsilon(t)\right),$$

donde $\upsilon(t) = \dfrac{\alpha(t)}{\gamma(t)}t.$

Demostración. El estimador lineal local $\widehat{S}(t)$ puede escribirse de la forma

$$\widehat{S}(t) = \int_0^\tau \overline{K}_{t,h}(t-s)\widetilde{S}(s)Y^{(n)}(s)ds,$$

donde $Y^{(n)}(s) = \sum_{i=1}^n Y_i(s)$.

Teniendo en cuenta los trabajos de Nielsen y Tanggaard (2001) y definiendo

$$S^*(t) = \int_0^\tau \overline{K}_{t,h}(t-s)S(s)Y^{(n)}(s)ds,$$

el error de estimación estocástico $\widehat{S}(t) - S(t)$ puede separarse en dos términos

$$\widehat{S}(t) - S(t) = (\widehat{S}(t) - S^*(t)) + (S^*(t) - S(t)) = V(t) + B(t),$$

$V(t)$ es el término aleatorio que converge en distribución a la normal, y $B(t)$ es el término estable que converge en probabilidad a una constante.

En primer lugar, se analiza $B(t)$ como sigue

$$B(t) = S^*(t) - S(t) = \int_0^\tau \overline{K}_{t,h}(t-s)S(s)Y^{(n)}(s)ds - S(t).$$

Puesto que $\int_0^\tau \overline{K}_{t,h}(t-s)Y^{(n)}(s)ds = 1$, la expresión anterior puede escribirse de la forma

$$B(t) = \int_0^\tau \overline{K}_{t,h}(t-s)\left(S(s) - S(t)\right)Y^{(n)}(s)ds.$$

Suponiendo que $S \in C^{(2)}([0,\tau])$, y mediante un desarrollo en serie de Taylor se tiene

$$B(t) = \int_0^\tau \overline{K}_{t,h}(t-s)\Bigg(S(t) + S'(t)(s-t) + S''(t)\frac{(s-t)^2}{2} + o_P\left((s-t)^2\right) - S(t)\Bigg)Y^{(n)}(s)ds.$$

Teniendo en cuenta que $\overline{K}_{t,h}$ puede interpretarse como un núcleo de segundo orden con respecto a la medida κ, donde $d\kappa(s) = Y^{(n)}(s)ds$ (ver Nielsen y Tanggaard (2001)), se tiene que $\int_0^\tau \overline{K}_{t,h}(t-s)(t-s)Y^{(n)}(s)ds = 0$.

Por tanto,

$$B(t) = \int_0^\tau \overline{K}_{t,h}(t-s)\left(S''(t)\frac{(s-t)^2}{2} + o_P\left((s-t)^2\right)\right)Y^{(n)}(s)ds.$$

Nielsen y Tanggaard (2001) demuestran que el núcleo lineal local estocástico $\overline{K}_{t,h}(t-s)$ es asintóticamente equivalente al núcleo $K_h(t-s)\{Y^{(n)}(s)\}^{-1}$, donde $K_h(t-s) = h^{-1}K\left((t-s)h^{-1}\right)$. Definiendo el momento de segundo orden como $\mu_2(K) = \int u^2K(u)du$, se concluye que

$$B(t) = \frac{1}{2}h^2S''(t)\mu_2(K) + o_P(h^2).$$

En segundo lugar, se analiza $V(t)$

$$V(t) = \widehat{S}(t) - S^*(t) = \int_0^\tau \overline{K}_{t,h}(t-s)\left(\widetilde{S}(s) - S(s)\right) Y^{(n)}(s)ds.$$

Se consideran las siguientes aproximaciones

$$\widetilde{S}(s) = \exp\left(-\int_0^s \frac{dN^{(n)}(u)}{Y^{(n)}(u)}\right) = \exp\left(-\widetilde{\Lambda}(s)\right) \approx 1 - \widetilde{\Lambda}(s),$$
$$S(s) = \exp\left(-\int_0^s \alpha(u)du\right) = \exp\left(-\Lambda(s)\right) \approx 1 - \Lambda(s),$$

Sustituyendo $\widetilde{S}(s)$ y $S(s)$ en la expresión $V(t)$ se tiene

$$\begin{aligned}
V(t) &= \int_0^\tau \overline{K}_{t,h}(t-s)\left[1 - \widetilde{\Lambda}(s) - (1-\Lambda(s))\right] Y^{(n)}(s)ds \\
&= \int_0^\tau \overline{K}_{t,h}(t-s)\left[-\widetilde{\Lambda}(s) + \Lambda(s)\right] Y^{(n)}(s)ds \\
&= -\int_0^\tau \overline{K}_{t,h}(t-s)\left[\int_0^s \left(\frac{dN^{(n)}(u)}{Y^{(n)}(u)} - \alpha(u)du\right)\right] Y^{(n)}(s)ds \\
&= -\int_0^\tau K_h(t-s)\left(\int_0^s \frac{dN^{(n)}(u) - \alpha(u)Y^{(u)}(u)du}{Y^{(u)}(u)}\right) ds \\
&= -\int_0^\tau K_h(t-s)\left(\int_0^s \frac{dM^{(n)}(u)}{Y^{(n)}(u)}\right) ds,
\end{aligned}$$

donde $M^{(n)}(t) = N^{(n)}(t) - \alpha(t)Y^{(n)}(t)$ es una martingala local de cuadrado integrable, y nuevamente se ha utilizado la equivalencia asintótica entre núcleos. Si se intercambia el orden de las integrales en la expresión anterior y se realiza un cambio de variable apropiado, se tiene que

$$\begin{aligned}
V(t) &= -\int_0^\tau \left(\int_u^\tau K_h(t-s)ds\right) \frac{dM^{(n)}(u)}{Y^{(n)}(u)} \\
&= \int_0^\tau \left(-\mathcal{K}\left(\frac{t-u}{h}\right)\right) \frac{dM^{(n)}(u)}{Y^{(n)}(u)},
\end{aligned}$$

donde $\mathcal{K}(x) = \int_{-\infty}^x K(\mathrm{v})d\mathrm{v}$, es una función predecible y por lo tanto se puede proceder siguiendo el teorema límite para martingalas dado en la Subsección (1.2.2).

Para obtener la distribución límite, basta con demostrar que para alguna sucesión $a_n \to \infty$ y para alguna función positiva $\upsilon(t)$, se tiene que

$$a_n \langle V \rangle (t) \xrightarrow[n\to\infty]{\mathcal{P}} \upsilon(t),$$

donde

$$\langle V \rangle (t) = \int_0^\tau \left(\mathcal{K}\left(\frac{t-u}{h} \right) \right)^2 \frac{1}{Y^{(n)}(u)} \alpha(u) du.$$

Bajo las hipótesis *(A.2)-(A.4)* del Teorema 4, donde γ es una función positiva tal que $\gamma \in C^{(1)}([0,\tau])$, se obtiene que la anterior integral puede aproximarse mediante la expresión

$$\frac{h\alpha(t)}{n\gamma(t)} \left(\int_{-\infty}^{t/h} \mathcal{K}^2(\mathrm{v}) d\mathrm{v} \right). \tag{2.25}$$

En la mayoría de los casos, la función núcleo K es una función de densidad simétrica con dominio en el intervalo $[-1,1]$, por lo que $\mathcal{K}$ indica la correspondiente función de distribución acumulada, y entonces la expresión (2.25) puede escribirse de la forma

$$\frac{h\alpha(t)}{n\gamma(t)} \left(\int_{-1}^{1} \mathcal{K}^2(\mathrm{v}) d\mathrm{v} + \int_1^{t/h} d\mathrm{v} \right) = \frac{h\alpha(t)}{n\gamma(t)} \left(R(\mathcal{K}) + \frac{t}{h} - 1 \right), \tag{2.26}$$

donde $R(\mathcal{K}) = \int_{-1}^{1} \mathcal{K}^2(\mathrm{v}) d\mathrm{v}$ es una constante que sólo depende del núcleo. Esta ecuación puede reescribirse como

$$\frac{\alpha(t)}{\gamma(t)} \left(\frac{R(\mathcal{K}) - 1}{nh^{-1}} + \frac{t}{n} \right). \tag{2.27}$$

El término principal de la expresión (2.27) es de orden $o_P(n^{-1})$, por lo que puede aproximarse mediante

$$\frac{\alpha(t)}{\gamma(t)} \frac{t}{n} + o_P\left(n^{-1}\right).$$

□

Hay que señalar que este resultado concuerda con la discusión del Capítulo 1 de Gámiz, Kulasekera, Limnios y Lindquist (2011), es decir, el orden de convergencia

de la varianza del estimador lineal local coincide con el proporcionado por la función de distribución acumulada empírica, $o_P\left(n^{-1}\right)$. Sin embargo, de (2.27), también se tiene que la varianza posee un término adicional de orden inferior, $o_P(hn^{-1})$, que sólo depende del núcleo y que puede ser negativo. En particular, para el núcleo de Epanechnikov, se obtiene que $R\left(\mathcal{K}\right) - 1 = -0{,}2571429$. Esto significa que el estimador lineal local presentado aquí proporciona una mejora en términos de eficiencia.

2.4. Estudio de simulación

En esta sección se evalúan las propiedades del procedimiento de estimación propuesto en la Sección 2.3 para muestras finitas.

2.4.1. Objetivos y especificaciones del estudio

En primer lugar, el estudio se centra en la parte paramétrica del modelo, esto es, se evalúa la precisión de los estimadores de los parámetros de regresión y se comparan con métodos alternativos. En segundo lugar, se evalúa la precisión del estimador lineal local de la distribución subyacente.

Para lograr el primer objetivo, se compara el modelo AFT y el modelo PH en un escenario donde se cumplen las hipótesis de ambos métodos y las estimaciones obtenidas de los parámetros de regresión son directamente comparables. Por lo tanto, en un primer estudio de comparación, no se tiene en cuenta la familia de distribuciones base, sino que se estudian las estimaciones de los parámetros de regresión que describen los efectos de las covariables sobre los tiempos de vida para el caso AFT, y la función de azar para el caso PH. Además, se mide el error de estimación obtenido mediante los dos ajustes semi-paramétricos y se comparan estos errores con el obtenido usando toda la información del modelo específico que se ha simulado.

Los cálculos computacionales han sido desarrollados utilizando el software es-

tadístico R Core Team (2015), concretamente, para implementar el modelo AFT paramétrico se han utilizado los libros *survival* y *eha*, mientras que el ajuste paramétrico ha sido implementado utilizando la función *aftreg* incluida en el ya citado libro *eha*, (ver Broström (2012)).

Dado que el modelo de regresión paramétrico de Weibull puede verse como un modelo de PH y como un modelo AFT, se decide utilizar esta familia de distribuciones para comparar la mejora de las dos aproximaciones. Se genera un modelo AFT con distribución base Weibull bajo diferentes esquemas de muestreo considerando muestras que presenten diferentes tasas de censura y truncamiento.

Desde la perspectiva del modelo PH, específicamente, se considera la siguiente expresión

$$\lambda(t; X) = \lambda_0(t) \exp\left(\phi_1 X_1 + \phi_2 X_2\right), \tag{2.28}$$

donde $\lambda_0(t)$ es una función de azar y $\boldsymbol{\phi} = (\phi_1, \phi_2)^\top$ es un vector de coeficientes de regresión asociado al vector de covariables $\mathbf{X} = (X_1, X_2)^\top$.

En términos del modelo AFT se tiene que el logaritmo del tiempo de vida de un sujeto con vector de covariables $\mathbf{X} = (X_1, X_2)^\top$ puede ser escrito como

$$\log T = \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \tag{2.29}$$

donde $\varepsilon = \log T_0$, siendo T_0 el tiempo de vida de un sujeto sin covariables, es decir, la población base, con $\beta_1 = -\phi_1$, y $\beta_2 = -\phi_2$.

Para llevar a cabo las simulaciones se considera un escenario similar al descrito en Orbe, Ferreira y Núñez-Antón (2002). Se asume que X_1 tiene distribución uniforme $\mathcal{U}(0, 2)$, X_2 también tiene distribución uniforme $\mathcal{U}(3, 9)$, y $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (1, 3)^\top$.

Para la población base, T_0, se considera una distribución de tipo Weibull con parámetro de escala $sh = 1$ y tres parámetros de forma diferentes, $sh = 5$, $sh = 0.5$ y $sh = 1$, con el objetivo de considerar funciones de azar crecientes, decrecientes y constantes como la distribución Exponencial, respectivamente. Bajo este marco, los modelos PH y AFT son directamente comparables dado que $\phi = -\beta$.

2.4.2. Generación de datos de supervivencia

Las muestras aleatorias se han generado de acuerdo al siguiente algoritmo:

1. Se genera una muestra de tiempos de fallo $\{T_1, T_2, \ldots, T_n\}$ de un modelo de regresión Exponencial con $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top = (1, 3)$, $X_1 \in \mathcal{U}(0, 2)$ y $X_2 \in \mathcal{U}(3, 9)$, siendo ε una distribución de valores extremos mínima estándar.

2. Se genera una muestra de tiempos censurados por la derecha independientes de los tiempos de fallo, $\{C_1, C_2, \ldots, C_n\}$, con una distribución Uniforme $\mathcal{U}(0, \xi)$, donde ξ es elegida apropiadamente para conseguir la tasa de censura deseada $p_c\,\%$ (10 %, 30 % y 45 %).

3. Se define el indicador de censura de la forma $\delta_i = I\,[T_i \leq C_i]$. Entonces, se representa la i-ésima observación mediante el par (Z_i, δ_i), con $Z_i = T_i \wedge C_i$, para cada $i = 1, 2, \ldots, n$.

4. Se generan n tiempos de truncamiento $\{L_1, L_2, \ldots, L_n\}$ independientes de la censura y de los tiempos de vida, a partir de una distribución Uniforme $\mathcal{U}\,(0, \varpi)$, donde ϖ es elegida apropiadamente para conseguir una tasa de truncamiento deseada, p_t= 10 %. Los sujetos para los cuales $L_i \leq Z_i$ se mantienen en la muestra y los demás son descartados.

Cada muestra simulada consiste en una secuencia de tripletas $\{(L_i, Z_i, \delta_i);\ i = 1, 2, \ldots, n^*\}$, donde $L_i \leq Z_i$ para cada $i = 1, 2, \ldots, n^* \leq n$. Se repite el procedimiento hasta $R = 2000$ veces considerando tamaños muestrales $n = 50, 100$ y 200, y distintas combinaciones de niveles de censura y truncamiento.

2.4.3. Criterios de error

A partir de una muestra simulada, por una parte, se estiman el sesgo y el error estándar de los estimadores de los parámetros de regresión utilizando el modelo AFT especificado en (2.29) y se comparan con los resultados obtenidos utilizando

el modelo de PH dado en (2.28). Por otra parte, se proporciona una medida del error de la estimación lineal local de la distribución de supervivencia base. En otras palabras, para las estimaciones de los parámetros de regresión ($\widehat{\beta}$ o $\widehat{\phi}$) se calcula una aproximación del valor del sesgo mediante la expresión $\text{Bias}[\widehat{\beta}] = E[\widehat{\beta}] - \beta$ y, la varianza $\text{Var}[\widehat{\beta}] = \text{se}^2[\widehat{\beta}] = E\left[\left(\widehat{\beta} - E[\widehat{\beta}]\right)^2\right]$. Finalmente, se aproxima el error cuadrático medio (MSE)

$$\text{MSE}(\widehat{\beta}) = \text{Var}[\widehat{\beta}] + \text{Bias}^2[\widehat{\beta}].$$

Con el fin de evaluar el funcionamiento del estimador lineal local de la supervivencia base, se considera el error cuadrático promedio (ASE). Para una muestra en particular de tamaño n se define como

$$\text{ASE} = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{S}_0\left(\exp\left(-\widehat{\beta}_1 x_{1,i} - \widehat{\beta}_2 x_{2,i}\right) t_i\right) - S_0\left(\exp\left(-\widehat{\beta}_1 x_{1,i} - \widehat{\beta}_2 x_{2,i}\right) t_i\right)\right]^2.$$

En primer lugar se estiman los parámetros de regresión como se explica en la Sección 2.3.1 y, en segundo lugar se obtiene la función de supervivencia evaluada en los datos de los tiempos observados y transformados a escala de tiempos base (ver Sección 2.3.3). Para calcular el estimador lineal local se considera la función núcleo de tipo Epanechnikov, $K(t) = \frac{3}{4}(1-t^2)I\left[\mid t \mid \leq 1\right]$, y el parámetro de suavizado es elegido utilizando una regla de referencia sugerida en Gámiz *et al.* (2011), p. 44.

2.4.4. Resultados y conclusiones

Los resultados de las simulaciones vienen resumidos en las Tablas 2.1-2.3 y en las Figuras 2.1-2.3. Se deduce que altos niveles de censura tienden a incrementar la varianza de los estimadores, por tanto, las estimas pierden precisión cuando las tasas de censura se incrementan. En términos del MSE, de acuerdo con las Tablas 2.1-2.3, el modelo AFT proporciona estimas más precisas que el modelo de PH en casi todos los casos. Para el modelo PH puede verse en la Tabla 2.2 que el resumen de la medida del error proporciona valores que son significativamente

mayores que los proporcionados por el modelo de AFT. Se han inspeccionado las muestras simuladas y se ha detectado que para pocas muestras, el valor del sesgo estimado es extremadamente grande, lo que da lugar a un error muy alto.

Tabla 2.1: Aproximación de Monte Carlo del MSE para los modelos PH, AFT y Exponencial.

Tamaño	Censura 10 %			Censura 30 %			Censura 45 %		
muestra	PH	AFT	Exp	PH	AFT	Exp	PH	AFT	Exp
$n = 50$	0.4750	0.1368	0.0975	0.7336	0.1763	0.1335	1.2125	0.2458	0.2122
$n = 100$	0.1715	0.0770	0.0461	0.2085	0.0772	0.0545	0.3116	0.1197	0.1330
$n = 200$	0.0689	0.0444	0.0205	0.0932	0.0503	0.0277	0.1403	0.0671	0.0393

Tabla 2.2: Aproximación de Monte Carlo del MSE para los modelos PH, AFT y Weibull(1, 5).

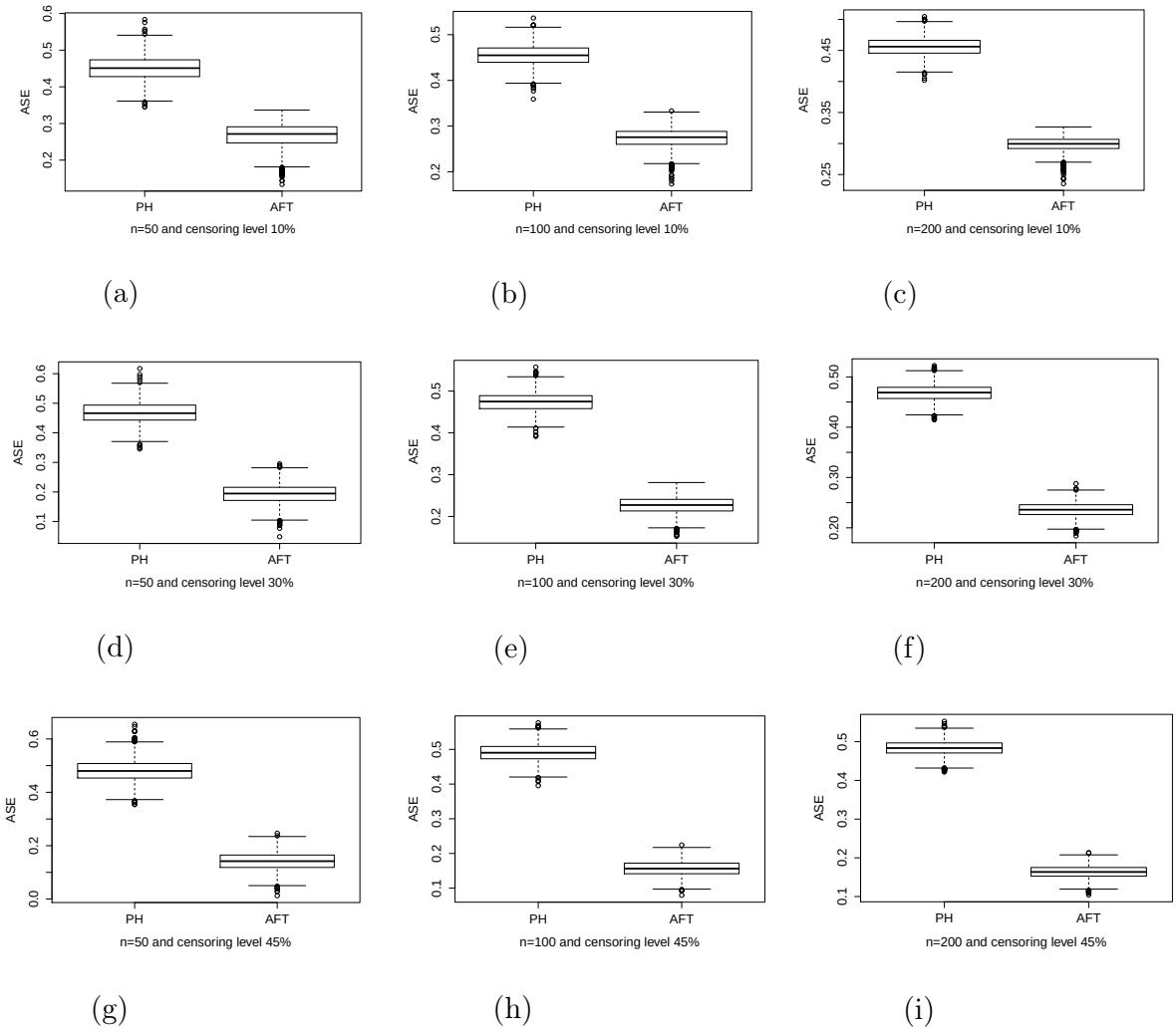
Tamaño	Censura 10 %			Censura 30 %			Censura 45 %		
muestra	PH	AFT	Weib	PH	AFT	Weib	PH	AFT	Weib
$n = 50$	0.6472	0.0048	0.0036	0.9302	0.0072	0.0057	1.3221	0.0108	0.0082
$n = 100$	0.2014	0.0028	0.0018	0.2282	0.0033	0.0022	0.3523	0.0048	0.0032
$n = 200$	0.0656	0.0016	0.0008	0.0852	0.0018	0.0010	0.1379	0.0027	0.0014

Tabla 2.3: Aproximación de Monte Carlo del MSE para los modelos PH, AFT y Weibull(1, 0.5).

Tamaño	Censura 10 %			Censura 30 %			Censura 45 %		
muestra	PH	AFT	Weib	PH	AFT	Weib	PH	AFT	Weib
$n = 50$	0.7062	0.5983	0.3682	1.1566	0.8357	0.5627	1.6963	0.9948	0.8012
$n = 100$	0.3223	0.3712	0.1828	0.4365	0.3959	0.2509	0.5625	0.5001	0.3247
$n = 200$	0.1384	0.2166	0.0840	0.1909	0.2309	0.1149	0.2676	0.3046	0.1595

Además, como es de esperar, los modelos semi-paramétricos son menos precisos cuando se estima un modelo paramétrico que utiliza la verdadera distribución de probabilidad de la que se han generado los datos.

Los valores del ASE del modelo AFT son significativamente más pequeños que los obtenidos mediante un modelo PH para todos los casos considerados. Hay que tener en cuenta que, como es de esperar, los valores del ASE decrecen conforme se va incrementando el tamaño muestral, ver Figuras 2.1-2.3.

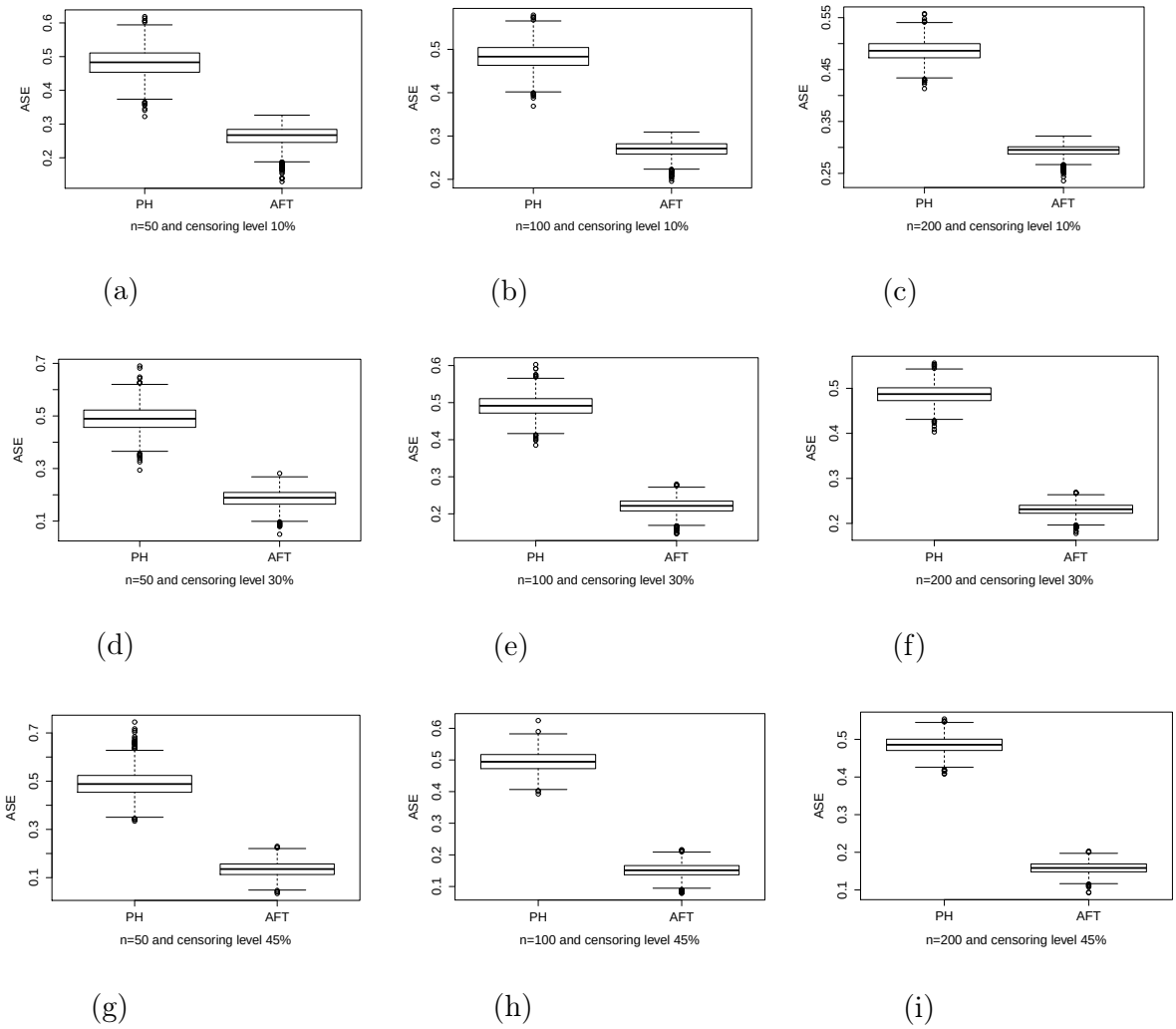


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figura 2.1: Caso 1. Modelo Exponencial. Box-plots de las estimaciones del ASE para los modelos PH y AFT con $n = 50$, 100 y 200, con nivel de truncamiento $p_t = 10\,\%$ y niveles de censura $p_c = 10\,\%$, $30\,\%$ y $45\,\%$.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figura 2.2: Caso 2. Modelo Weibull con parámetro de forma igual a 5. Box-plots de las estimaciones del ASE para los modelos PH y AFT con $n = 50$, 100 y 200, con nivel de truncamiento $p_t = 10\,\%$ y niveles de censura $p_c = 10\,\%$, $30\,\%$ y $45\,\%$.

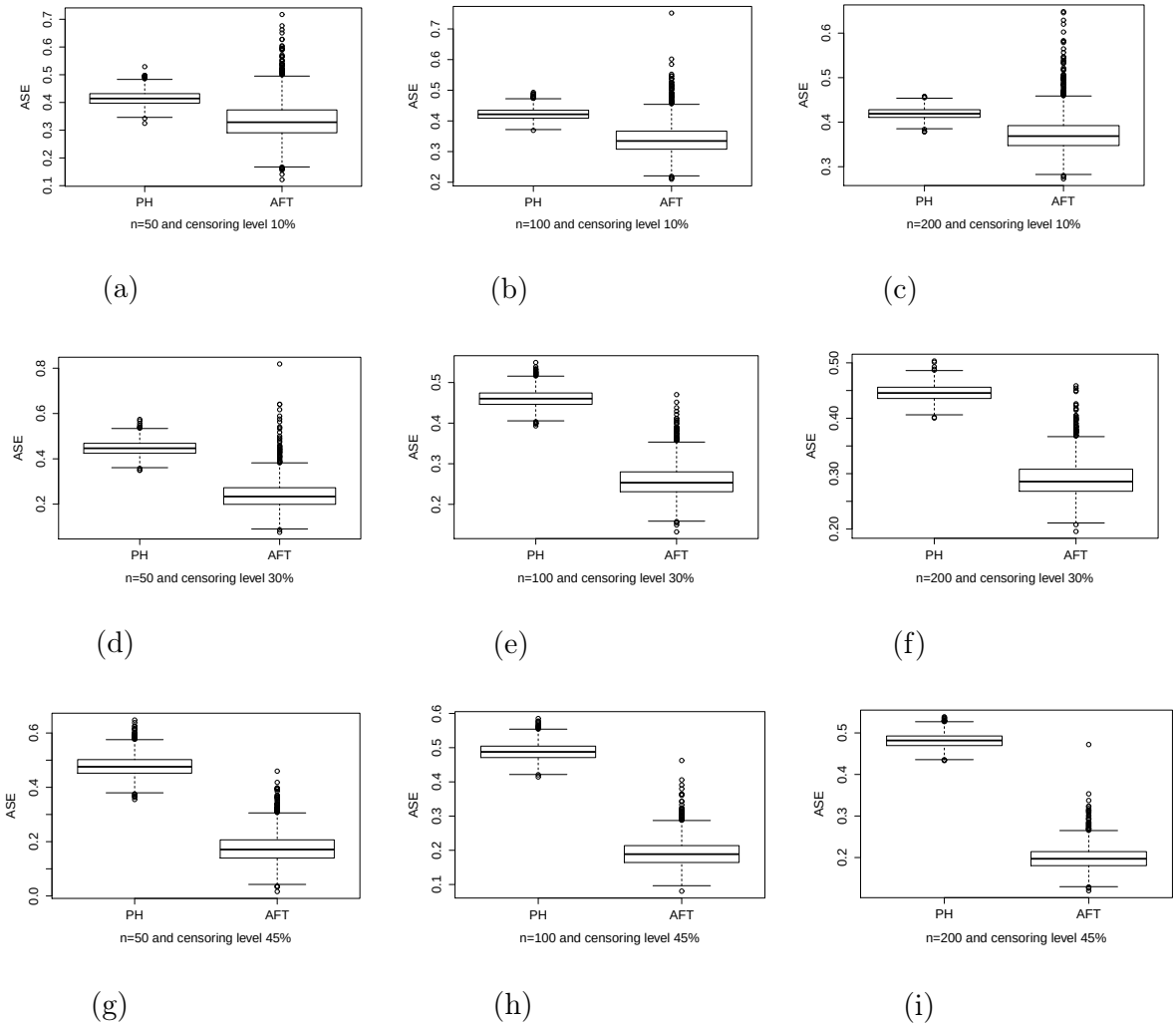


Figura 2.3: Caso 3. Modelo Weibull con parámetro de forma igual a 0.5. Box-plots de las estimaciones del ASE para los modelos PH y AFT con $n = 50$, 100 y 200, con nivel de truncamiento $p_t = 10\,\%$ y niveles de censura $p_c = 10\,\%$, $30\,\%$ y $45\,\%$.

2.5. Aplicación con datos reales

2.5.1. Presentación de los datos y objetivos del análisis

En esta sección se analizan los datos de fallo registrados en un sistema de suministro de agua de una ciudad de tamaño medio de la costa mediterránea española. El objetivo es evaluar el tiempo de vida de las tuberías utilizando la aproximación semi-paramétrica descrita anteriormente. Se considera el mismo conjunto de datos utilizado por Carrión, Solano, Gámiz y Debón (2010), donde se evalúa la probabilidad de fallo de las tuberías utilizando un modelo de riesgos proporcionales de Cox. El conjunto de datos contiene 26.034 registros de secciones de tuberías con información relativa al año de instalación de las tuberías, la longitud de cada tubería, el diámetro de la sección de cada tubería, el material de cada tubería, las condiciones de tráfico registrado para cada tubería así como su tiempo de fallo (en años). Según el conjunto de datos, se han utilizado cuatro tipos diferentes de *material* para construir dichas tuberías: fundición dúctil, fundición gris, polietileno y fibrocemento. También se han considerado tres tipos de *condiciones del tráfico* relativas al área en la que están sometidas las tuberías: acera, tráfico normal y tráfico pesado.

En el análisis no se considera el hecho de que una sección de tubería pueda fallar más de una vez y se restringe la observación a una ventana de intervalos de tiempo comprendida entre los años 2000-2005, ya que a partir de este año es cuando se empiezan a utilizar en este contexto los Sistemas de Información Geográfica (SIG). Hay que tener en cuenta que este esquema de muestreo induce un truncamiento por la izquierda así como una censura por la derecha, destacando que los datos presentan una gran tasa de censura superior al 98 %.

Se define la v.a. $L = \max\{0, 2000 - \text{fecha del año de la instalación}\}$ para representar los tiempos de truncamiento por la izquierda y se elige el valor 0 si el elemento correspondiente no está truncado por la izquierda. El tiempo de fallo se representa por la v.a. $T =$ fecha del fallo registrado $-$ fecha del año de instalación y el tiempo

de censura mediante la variable $C = 2006 -$ fecha de la instalación de la tubería. Todos los datos vienen registrados en años. Se supone que L, T y C son mutuamente independientes, no-negativas y que el esquema de censura es no-informativo, donde G, F y H son las funciones de distribución de L, T y C, respectivamente.

En primer lugar, se propone un modelo de PH para analizar el periodo de tiempo hasta que sucede el fallo en una tubería en particular de la red de suministro de agua considerada. Se define un vector con las siguientes covariables:

dúctil: Tipo de material (1=fundición dúctil, 0=otros).

gris: Tipo de material (1=fundición gris, 0=otros).

polietileno: Tipo de material (1=polietileno, 0=otros).

acera: Nivel del tráfico rodado (1=acera, 0=otros).

normal: Nivel del tráfico rodado (1=normal, 0=otros).

longitud: La longitud de la tubería en (m).

diámetro: El diámetro de la sección de la tubería en (mm).

Hay que tener en cuenta que para este modelo las categorías de referencia que se han considerado son el nivel de la covariable material *fibrocemento* ya en desuso debido a los efectos cancerígenos que puede producir, y el nivel de la covariable tráfico *tráfico pesado*.

Bajo las especificaciones de muestreo arriba indicadas, el objetivo es estimar el tiempo de vida de las tuberías que transportan agua hasta la rotura de la misma. Se comienza ajustando un modelo de PH de Cox para lo cual previamente se comprueba si se cumple la hipótesis de PH de dicho modelo.

Tabla 2.4: Resultados del test sobre la hipótesis de riesgos proporcionales.

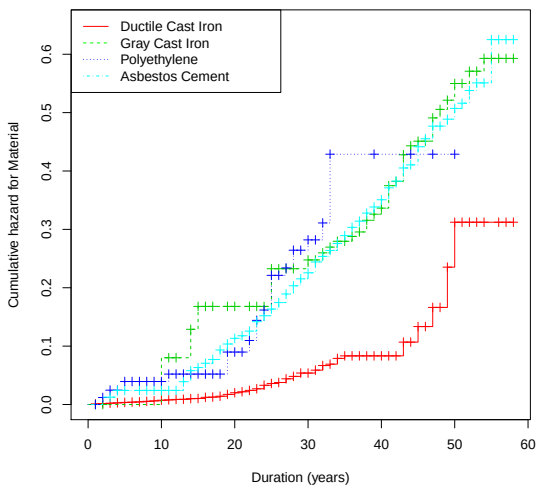
Covariable	ϱ	χ^2	p-valor
dúctil	0.1110	10.5	0.0012
gris	0.0007	0.0003	0.9859
polietileno	0.0396	0.983	0.3214
acera	0.1206	8.86	0.0029
normal	0.1388	11.8	0.0005
longitud	0.0395	0.529	0.4670
diámetro	0.0583	1.88	0.1707
Global		25.7	0.0005

2.5.2. Verificación de la hipótesis de riesgos proporcionales

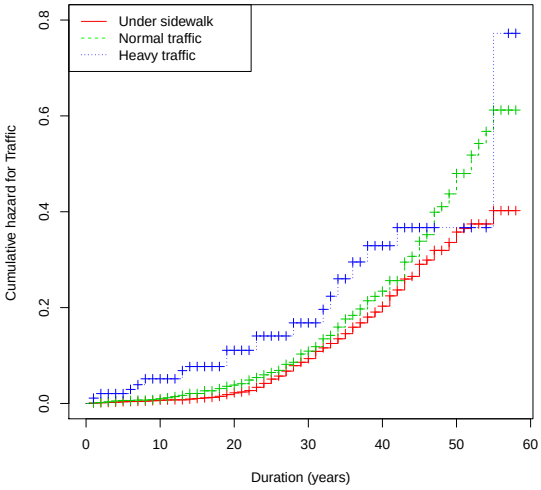
La hipótesis principal del modelo PH es la proporcionalidad de riesgos. Si dicha hipótesis se cumple, los logaritmos de las funciones de azar acumuladas deben describir curvas paralelas. Para ajustar el modelo PH a los datos se utiliza la función *cox.zph* que viene incluida en el libro *survival* de R, ver Therneau y Grambsch (2000). Los test de diagnóstico gráfico del modelo PH se basan en los residuos escalados de Schoendfeld, ver Sección 2.2.1 y Therneau y Lumley (2008).

Una inspección visual de la Figura 2.4 revela que las curvas correspondientes a las covariables *material*, *tráfico* y *diámetro* no son paralelas. A consecuencia de ello, el modelo PH es cuestionable en este estudio. Los resultados del test de los residuos de Schoendfeld para verificar la hipótesis de riesgos proporcionales, se muestran en la Tabla 2.4. Si el p-valor obtenido es menor que el nivel de significación 0.05, se concluye que la hipótesis de riesgos proporcionales no se cumple. El p-valor tan pequeño del test global sugiere que la hipótesis del modelo no se cumple en algún caso, concretamente habría que estudiar las covariables *material* y *tráfico*, como sugiere la tabla. Además, un análisis visual de la Figura 2.5 corrobora que tales covariables no cumplen la hipótesis de riesgos proporcionales.

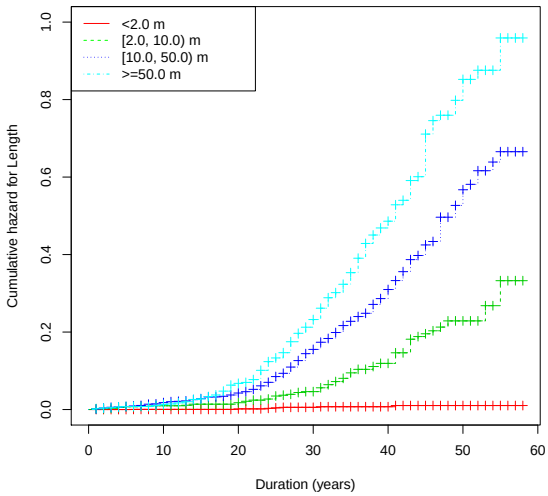
En este contexto, una manera de resolver el problema de la no-proporcionalidad de riesgos sería considerar covariables dependientes del tiempo en la formulación



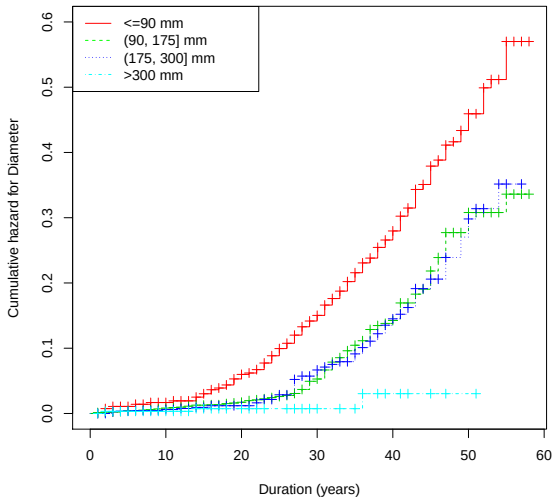
(a)



(b)

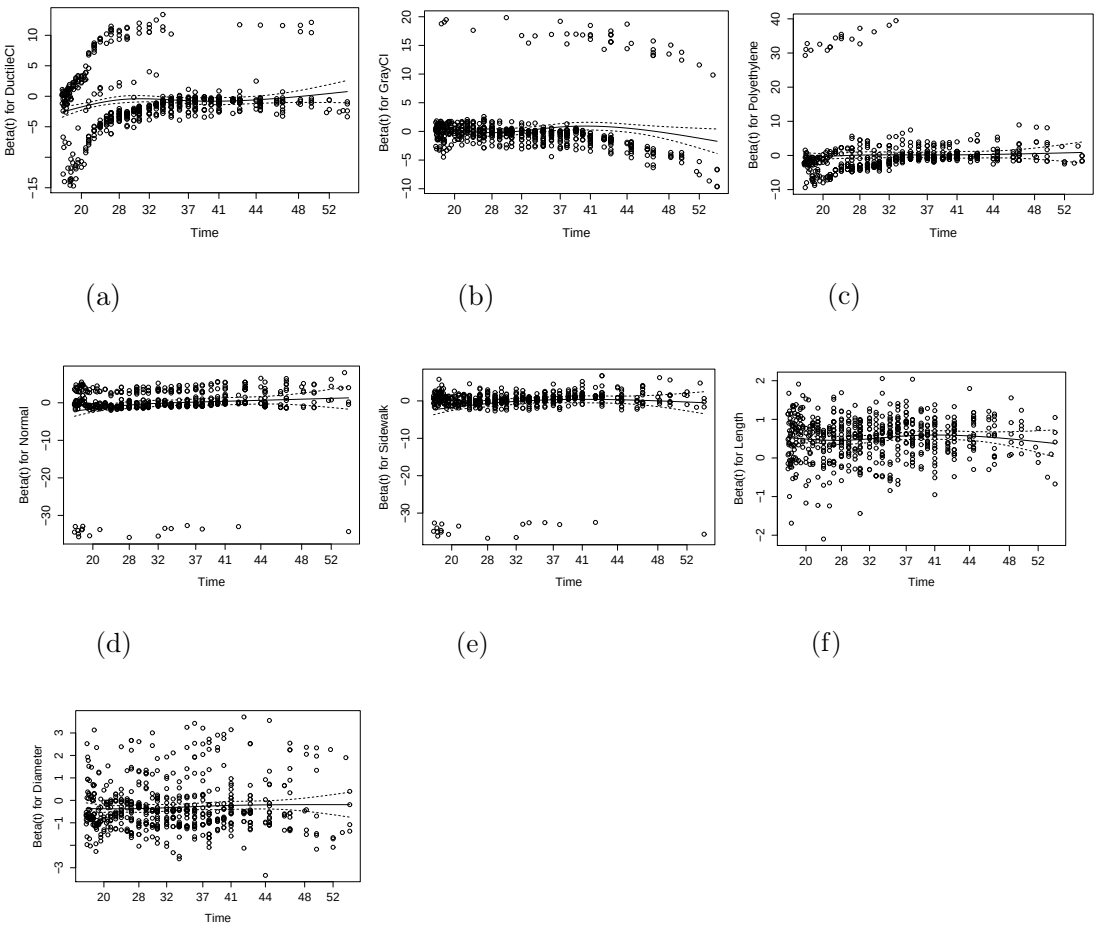


(c)



(d)

Figura 2.4: Funciones de azar acumulado estratificadas por las covariables material, tráfico, longitud y diámetro.



(a) (b) (c)

(d) (e) (f)

(g)

Figura 2.5: Gráfico de los residuos escalados de Schoenfeld para las covariables material, tráfico, longitud y diámetro.

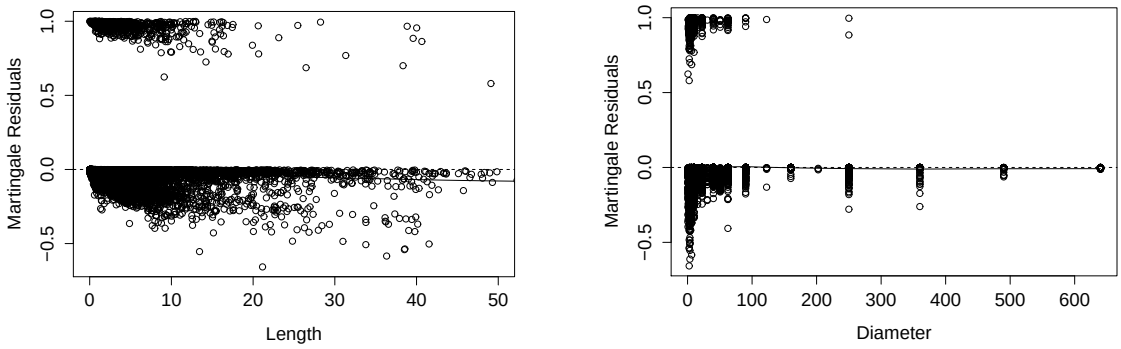
del modelo de regresión. Como alternativa, también puede resultar adecuado considerar un modelo de Cox estratificado según los niveles de las covariables *material* y/o *tráfico*, estrategia llevada a cabo por Carrion *et al.* (2010). Sin embargo, esta extensión del modelo de Cox presenta algunos inconvenientes como la pérdida de capacidad al cuantificar los efectos de estas covariables estratificadas que, en algunos casos, puede ser de gran interés como sucede en el conjunto de datos para las covariables *material* y *tráfico*.

Un aspecto adicional que cabe explorar es la no-linealidad del modelo, es decir, una forma funcional específica incorrecta de la parte paramétrica del modelo. Para detectar esto, se utilizan las gráficas de los residuos de martingala frente a las covariables. En este caso, en la Figura 2.6 se examinan los residuos de martingala frente a las covariables *longitud* y *diámetro* ya que la no-linealidad no es interpretable en el caso de las covariables dicotómicas. Al igual que en las gráficas de residuos de Schoenfeld, el suavizado también es útil para tener una idea clara de la gráfica. Las curvas suavizadas representadas en esta figura definidas mediante regresión lineal local (utilizando la función *lowess* de R), sugieren que la hipótesis de linealidad es razonable.

2.5.3. Ajuste del modelo AFT

Según el estudio de la Sección 2.5.2, la hipótesis de riesgos proporcionales no es adecuada para los datos de este estudio, por lo que se propone un procedimiento alternativo. En esta sección, se describen los resultados del modelo AFT semiparamétrico propuesto en la sección anterior. La estimación se lleva a cabo en dos pasos. En el primer paso, se calculan las estimaciones de los coeficientes de regresión en el modelo (2.17) sin suponer ninguna familia paramétrica para la distribución subyacente a los tiempos de vida. El segundo paso consiste en una estimación no-paramétrica de la función de supervivencia base.

Los coeficientes estimados junto a sus correspondientes errores estándar vienen



(a) (b)

Figura 2.6: Gráfico de los residuos de martingala para las covariables longitud y diámetro.

Tabla 2.5: Estimación de los coeficientes de regresión del modelo semi-paramétrico AFT.

Covariable	$\widehat{\beta}$	$\exp(\widehat{\beta})$	$\text{se}(\widehat{\beta})$	I.C. Bootstrap al 95 %
dúctil	-0.3530	0.7025	0.0974	(-0.5439 , -0.1621)
gris	0.4134	1.5120	0.1058	(0.2059 , 0.6209)
polietileno	-0.7449	0.4747	0.1838	(-1.1052 , -0.3845)
acera	1.2392	3.4529	0.5466	(0.1677 , 2.3107)
normal	1.5409	4.6690	0.5058	(0.5495 , 2.5323)
longitud	0.2594	1.2962	0.0727	(0.1169 , 0.4019)
diámetro	-0.2725	0.7614	0.0619	(-0.3939 , -0.1511)

dados en la Tabla 2.5. Los errores estándar se han calculado utilizando el método bootstrap sugerido por Gross y Lai (1996) en el que se han considerado B=10000 muestras bootstrap.

Los coeficientes de regresión del modelo AFT vienen dados de la forma $\exp(\beta)$, y se interpretan, para cada covariable, de la siguiente forma:

fundición dúctil: el tiempo de fallo correspondiente a una tubería hecha de fundición dúctil disminuye en un 29.75 % con respecto a los otros.

fundición gris: el tiempo de fallo correspondiente a una tubería hecha de fundición gris aumenta en un 51.20 % con respecto a los otros.

polietileno: el tiempo de fallo correspondiente a una tubería hecha de polietileno disminuye en un 52.53 % con respecto a los otros.

acera: el tiempo de fallo correspondiente a una tubería sometida a un tráfico de acera aumenta en un 245.29 % con respecto a los otros.

normal: el tiempo de fallo correspondiente a una tubería sometida a un tráfico normal aumenta en un 366.90 % con respecto a los otros.

longitud: el aumento en 1 m de longitud de tubería corresponde a un aumento del tiempo de fallo del 29.62 %.

diámetro: el aumento en 1 mm en el diámetro de la sección de la tubería corresponde a una disminución del tiempo de fallo del 23.86 %.

Para las covariables numéricas *longitud* y *diámetro*, se considera la media de los valores como niveles de referencia. Para estimar la función de supervivencia base se considera el estimador lineal local (2.24). Al igual que en la Sección 2.4 se calcula dicho estimador utilizando la función núcleo de Epanechnikov con un parámetro de suavizado cuyo valor se elige según una regla de referencia (siguiendo las sugerencias de Gámiz *et al.* (2011), p. 44), que proporciona el valor $h = 0.0707$. El estimador lineal local resultante de la función de supervivencia base se muestra en la Figura 2.7.

Como conclusiones puede decirse que cuando la hipótesis de riesgos proporcionales es dudosa, los resultados obtenidos del modelo PH de Cox pueden llevar a conclusiones erróneas. Por el contrario, el modelo AFT semi-paramétrico proporciona resultados más adecuados.

El hecho de asumir hipótesis paramétricas inadecuadas puede llevar a conclusiones erróneas derivadas de una mala especificación de la distribución de probabilidad

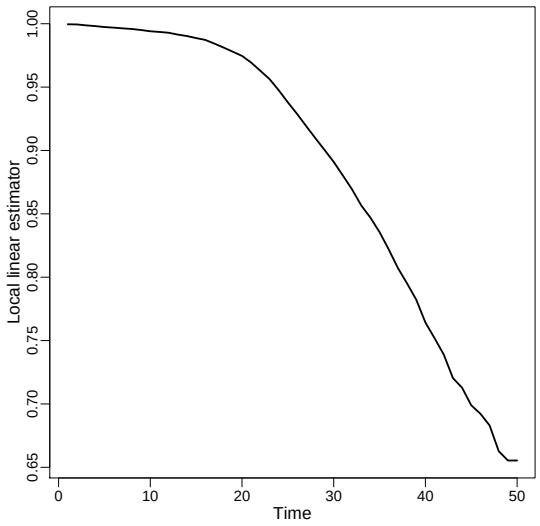


Figura 2.7: Estimador lineal local de la función de supervivencia base.

que subyace a los tiempos de fallo. Por tanto, si la distribución de probabilidad es desconocida (que es lo que sucede en la mayoría de las situaciones prácticas), resulta más conveniente considerar un enfoque no paramétrico.

Los métodos descritos en este capítulo permiten evaluar la influencia que una serie de covariables tiene sobre el tiempo de vida, tal y como se ilustra en el caso de las tuberías analizadas en el sistema de suministro de agua presentado en la sección anterior. De hecho, el análisis descrito anteriormente muestra que las tuberías menos propensas al fallo tienen las siguientes características: menores longitudes, menores diámetros, hechas con materiales de fundición dúctil o polietileno y situadas bajo un tráfico de acera y normal. Los resultados concuerdan con los derivados de trabajos previos que analizaron un conjunto de datos similar con un modelo diferente (ver por ejemplo Carrion *et al.* (2010)).

2.6. Conclusiones

En este capítulo se ha propuesto una estrategia novedosa que permite estimar un modelo AFT bajo un enfoque no paramétrico. Como demuestra el estudio de simulación, aún cuando la hipótesis de riesgos proporcionales del modelo de Cox se cumple, el modelo AFT propuesto se comporta mejor en términos del error cuadrático medio. En el caso en el que la hipótesis de PH no se cumpla, el modelo semi-paramétrico de AFT ofrece una alternativa al modelo tradicional de Cox, como sucede en la aplicación práctica de la Sección 2.5. Por tanto, se utiliza el procedimiento semi-paramétrico para estudiar la influencia de algunas variables en la fiabilidad de las tuberías de una red de suministro de agua.

Además, en el estudio de simulación desarrollado, se compara el modelo semi-paramétrico AFT propuesto con la versión paramétrica del mismo, esto es, con un modelo paramétrico AFT con la verdadera distribución de probabilidad de los errores. Como es de esperar, la metodología semi-paramétrica propuesta es menos precisa, aunque esta pérdida de precisión es muy pequeña, por lo que puede compensarse ya que en la aproximación semi-paramétrica no se necesita hacer ninguna hipótesis paramétrica sobre el tiempo de fallo subyacente.

Además, se han comparado las funciones de supervivencia base para ambos modelos AFT y PH, y se ha confirmado que el modelo AFT es superior en términos de precisión.

Las ideas proporcionadas en este capítulo pueden ser muy valiosas cuando se toman decisiones relativas al diseño y construcción de redes de suministro de agua. Si bien, las conclusiones del análisis de datos de este trabajo son válidas no solo para el conjunto de datos estudiado aquí, sino también (hasta cierto punto) para cualquier otra red de suministro de agua.

Capítulo 3

Inferencia en un espacio de localización y escala de intensidades en PPNH

3.1. Introducción y objetivos

El análisis de datos con eventos recurrentes se utiliza en diversas áreas como pueden ser fiabilidad, medicina, ciencias sociales, economía, finanzas y criminología, en las que el analista está interesado en crear modelos estadísticos con el objetivo de modelizar el número de ocurrencias de eventos a lo largo del tiempo.

Tradicionalmente, uno de los modelos más utilizados para tratar datos de eventos recurrentes a lo largo del tiempo es el proceso de Poisson no homogéneo (PPNH). La magnitud que describe el número de eventos por unidad de tiempo es la función de intensidad (condicionada). Dicha función proporciona un patrón de ocurrencias en función del tiempo y se calcula en términos de la primera derivada del número esperado de ocurrencias, condicionado a la historia del proceso (ver Capítulo 1).

Como se indica en el trabajo de Krivtsov (2007), una inmensa mayoría de publicaciones sobre PPNHs considera sólo dos formas monótonas de la función de intensidad, el modelo log-lineal y el modelo *power law*. En esta memoria se supone

que la intensidad del PPNH coincide formalmente con la función de azar de la distribución de los tiempos de vida subyacentes. Por tanto, para modelizar la intensidad de un proceso PPNH se pueden utilizar formas paramétricas tradicionales de las funciones de azar de las distribuciones de tiempos de vida, o combinaciones de las mismas.

Estimar la intensidad es de gran utilidad para optimizar el rendimiento de los activos bajo supervisión. Con este objetivo, un análisis estadístico con hipótesis paramétricas incorrectas sobre el tiempo de vida subyacente, podría llevar a conclusiones erróneas. Por otra parte, en muchos casos no se necesita ninguna descripción detallada de la relación evento-tiempo, más bien, lo que el analista necesita saber es una idea sobre los cambios de tendencia críticos en la función de intensidad.

Teniendo en cuenta todas las consideraciones anteriores, el objetivo de este capítulo es proporcionar una herramienta gráfica eficaz para explorar las características subyacentes de la intensidad de PPNH, con el fin de detectar patrones constantes o monótonos, posibles cambios de tendencia, o en general, picos y valles a lo largo del tiempo. Con esta finalidad se ha desarrollado una extensión de la herramienta gráfica SiZer Map introducida por Chaudhuri y Marron (1999).

SiZer Map, abreviatura de *"significant zero crossing of the derivatives"*, es una poderosa herramienta gráfica exploratoria que fue introducida en un principio para funciones de densidad y de regresión. Esta herramienta utiliza un suavizado de tipo núcleo para estimar la estructura que subyace a los datos, y juega con el parámetro de suavizado como un parámetro de escala, para visualizar las características subyacentes en la función objeto de estudio. Las características que "realmente están ahí", es decir, las que no son un mero artefacto de la variabilidad muestral, se descubren a través de la construcción de intervalos de confianza para la primera derivada de la función. SiZer Map toma las ideas del enfoque de espacio y escala de Lindeberg (1994), que ha sido adoptado en los últimos años para diversos pro-

blemas de investigación y aplicaciones como puede verse en Godtliebsen, Marron y Chaudhuri (2002), Marron y de Uña-Álvarez (2004), Li y Marron (2005), Hanning y Lee (2006), González-Manteiga (2008), Martínez-Miranda, Raya-Miranda, González-Manteiga y González-Carmona (2008), Park y Kang (2008), Park, Hannig y Kang (2009, 2014), Park, Lee y Hannig (2010), Park, Hernández-Campos, Le, Marron, Park, Pipiras, Smith, Smith, Trovero y Zhu (2011), Oliveira, Crujeiras y Rodríguez-Casal (2013), entre otros. En este capítulo, se muestra el análisis de la extensión de la herramienta SiZer para la función de intensidad en un PPNH.

La estructura de este capítulo es similar a la del trabajo sometido Gámiz-Pérez, López-Montoya, Martínez-Miranda y Raya-Miranda (2017), si bien, en este capítulo se proporcionan más detalles y se hace un análisis más profundo de los ejemplos prácticos y las simulaciones. Concretamente, en la Sección 3.2 se presentan seis ejemplos reales que pueden ajustarse de manera adecuada a un PPNH. En la Sección 3.3 se describe un estimador de tipo núcleo lineal local para la función de intensidad y sus derivadas, así como una estimación de la varianza de estos estimadores. El estimador lineal local propuesto aquí está íntimamente relacionado con el estimador propuesto por Chen *et al.* (2008) y Chen *et al.* (2011) para funciones de intensidad en procesos de recuento. En la Sección 3.4 se presenta la herramienta SiZer propuesta por Chaudhuri y Marron (1999) en el contexto de las funciones de densidad. A continuación se propone una extensión para la función de intensidad y se definen los elementos necesarios para su construcción en este caso. En concreto se proponen intervalos de confianza para la derivada de la función intensidad, que permiten hacer inferencia sobre los cambios de tendencia en dicha función. Se describen dos métodos alternativos para construir los intervalos de confianza, uno basado en una aproximación Normal y el otro basado en un método consistente de tipo bootstrap sugerido por Cowling, Hall y Phillips (1996). Como se indicó antes, se consideran ambos métodos para definir intervalos de confianza de tipo puntuales y simultáneos. Con el objetivo de demostrar la eficiencia de la extensión de SiZer

propuesta para la intensidad, en la Sección 3.5 se realiza un estudio de simulación. En la Sección 3.6 se describen seis aplicaciones a datos reales, dos de ellas relativas a análisis sísmicos, otras dos relacionadas con el análisis de Fiabilidad, una relativa al análisis de temporales en el mar Ártico y otra sobre la toma de decisiones en emergencias marítimas. Finalmente, en la Sección 3.7 se muestran algunos resultados y conclusiones generales del capítulo así como futuras líneas de investigación en el tema.

3.2. Ejemplos de datos reales modelizados mediante un PPNH

A continuación se presentan seis ejemplos de datos con tiempos de eventos que se ajustan a un PPNH. (Como se verán en la Sección 3.6)

1. *Análisis de ocurrencia de réplicas seguidas a un terremoto de gran magnitud.* Se consideran dos conjuntos de datos. El primero consta de 2305 tiempos (en días) de la ocurrencia de réplicas que van sucediendo días después del terremoto principal ocurrido en Miyagi-Ken (Japón) el 26 de julio del año 2003 con una magnitud de 6.2 en la escala de Richter. Los datos están disponibles en el libro SAPP de R. El segundo consta de 150 tiempos de observación de réplicas (en días) después del terremoto principal ocurrido en Sichuan (China) el 12 de mayo del año 2008 con una magnitud de 7.9 en la escala de Richter, estos datos pueden encontrarse en la United States Geological Survey (USGS) en `https://es.wikipedia.org/wiki/Terremoto_de_Sichuan_de_2008`. La ocurrencia de réplicas puede considerarse como un proceso puntual en el tiempo, concretamente, como un PPNH. En la Sección 3.6.1 se explican más detalles de ambos ejemplos.

2. *Sistema hidráulico de máquinas de carga, acarreo y descarga (LHD).* En este ejemplo se describen unos datos de Fiabilidad. Los datos han sido tomados

de Kumar y Klefsjö (1992) consisten en tiempos entre fallos sucesivos (en horas, excluyendo reparaciones o tiempos de paradas) de sistemas hidráulicos de seis máquinas de LHD. El tiempo de fallo de cada máquina puede ser adecuadamente representado por un PPNH. Dado que las máquinas operan de forma independiente a lo largo del tiempo, si todos los fallos ocurriesen en todas las máquinas, este caso se podría considerar como un proceso general de fallo y este esquema sería un PPNH. En la Sección 3.6.2 se explican más detalles de este ejemplo.

3. *Sistema del tren de potencia de un autobús urbano*. En este ejemplo se describen nuevamente unos datos de Fiabilidad. Los datos han sido tomados de Guida y Pulcini (2009) consisten en 55 tiempos de fallo (medidos en kilómetros) del sistema del tren de potencia de un autobús urbano en los periodos comprendidos entre 1999 y 2004. Los fallos sucedidos en el tren de potencia del autobús están sujetos a reparación mínima, y el proceso puede representarse mediante un PPNH. En la Sección 3.6.3 se explica este ejemplo con más detalle.

4. *Ocurrencia de tormentas en el mar Ártico*. En este ejemplo se describen unos datos relativos al número de tormentas sucedidas en el mar Ártico durante un periodo de tiempo determinado. Los datos han sido proporcionados por Lee, Wilson y Crawford (1991), y consisten en 302 tiempos de ocurrencia de tormentas en meses. La ocurrencia de tormentas en el mar Ártico es un proceso puntual en el tiempo que puede considerarse como un PPNH. En la Sección 3.6.4 se explican más detalles de este ejemplo.

5. *Emergencias relativas a inmigración ilegal asistidas por Salvamento Marítimo en el sureste peninsular*. Los datos constan de 291 tiempos de ocurrencia (en días) relativos al rescate de pateras ocurridas en el sureste de España, en los periodos comprendidos entre los años 2012 y 2015. Los datos han sido

proporcionados por el Centro de Coordinación de Salvamento de Almería. La ocurrencia de la llegada de pateras a la costa española es un proceso puntual en el tiempo que puede considerarse como un PPNH. En la Sección 3.6.5 se explican más detalles de este ejemplo.

6. *Análisis de ocurrencia de accidentes en minas de carbón.* En este ejemplo se describen unos datos clásicos relativos a accidentes de industrias mineras ocurridos en el Reino Unido. Los datos son tomados de Jarrett (1984) y consisten en 191 tiempos (en días) en los que ha sucedido un accidente (explosión con más de 10 muertos) en minas de carbón en los periodos comprendidos entre el 15 de marzo de 1851 y el 22 de marzo de 1962. La ocurrencia de accidentes es un proceso puntual en el tiempo y por tanto puede considerarse como un PPNH. En la Sección 3.6.6 se describe con detalle este ejemplo.

3.3. Estimador lineal local de la función de intensidad de un PPNH

En esta sección se sugiere un estimador lineal local para la función de intensidad λ de un PPNH $\{N(t),\, t \geq 0\}$. Este estimador se obtiene utilizando una formulación discreta y el criterio de mínimos cuadrados locales sugerido por Nielsen y Tanggaard (2001). Este resultado puede verse como una versión discreta del estimador polinomial local propuesto por Chen *et al.* (2011), que hereda la intuición y la simplicidad del procedimiento de estimación lineal local en regresión y permite reducir el coste computacional de la aproximación con datos continuos. Bajo una formulación de tipo discreta, se supone que los datos son frecuencias o conteos sucesivos de ocurrencias del proceso en instantes de tiempo discretos, $\{x_0 = 0,\, x_1, x_2, \ldots, x_M\}$. Entonces, la información que se registra es de la forma $N_r = N(x_r) - N(x_{r-1})$, para $r = 1, 2, \ldots, M$. Puesto que el proceso es de tipo PPNH, N_r tiene una distribución de Poisson con parámetro $\mu_r = \Lambda(x_r) - \Lambda(x_{r-1})$, para $r = 1, 2, \ldots, M$, donde Λ

es la función de intensidad acumulada, $\Lambda(t) = \int_0^t \lambda(u)du$. Bajo la condición de suavidad supuesta se considera la aproximación $\mu_r \approx \lambda(x_r)(x_r - x_{r-1}) = \lambda_r \Delta_r$, con $\lambda_r = \lambda(x_r)$ y $\Delta_r = x_r - x_{r-1}$ con $r = 1, 2, \ldots, M$.

En esta situación, se define un estimador empírico de λ_r de la forma

$$\widehat{\lambda}_r = \frac{N_r}{\Delta_r}.$$

Bajo la aproximación lineal local se supone que, para cada x_r, la función de intensidad λ puede aproximarse localmente (en un entorno de x_r) por una función lineal, es decir, $\forall x \in (x_r - h, x_r + h)$, $\lambda(x) = \theta_0 + \theta_1(x - x_r)$, para un parámetro de suavizado h apropiado. Por tanto, en el punto de la rejilla x_r con $r = 1, 2, \ldots, M$, una estimación del término independiente θ_0 proporciona una estimación de $\lambda(x_r)$, mientras que una estimación de la pendiente θ_1 proporciona una estimación de la derivada $\lambda'(x_r)$.

Con este objetivo, se formula el siguiente problema de estimación de mínimos cuadrados

$$\begin{pmatrix} \widehat{\theta}_0 \\ \widehat{\theta}_1 \end{pmatrix} = \underset{\theta_0,\theta_1}{\operatorname{argmin}} \left\{ \sum_{m=0}^{M} \left[\widehat{\lambda}_m - \theta_0 - \theta_1 (x_m - x_r) \right]^2 K_h(x_m - x_r) \right\}.$$

Para obtener un estimador de la función de intensidad y de su derivada, es decir,

$$\begin{pmatrix} \widehat{\lambda}(x_r) \\ \widehat{\lambda}'(x_r) \end{pmatrix} = \begin{pmatrix} \widehat{\theta}_0 \\ \widehat{\theta}_1 \end{pmatrix},$$

se deriva $\sum_{m=0}^{M} \left[\widehat{\lambda}_m - \theta_0 - \theta_1 (x_m - x_r) \right]^2 K_h(x_m - x_r)$ respecto de los parámetros y se iguala a cero, obteniéndose que

$$-2 \sum_{m=0}^{M} \left[\widehat{\lambda}_m - \theta_0 - \theta_1 (x_m - x_r) \right] K_h (x_m - x_r) = 0,$$
$$-2 \sum_{m=0}^{M} \left[\widehat{\lambda}_m - \theta_0 - \theta_1 (x_m - x_r) \right] (x_m - x_r) K_h (x_m - x_r) = 0,$$

es decir,

$$\sum_{m=0}^{M} \widehat{\lambda}_m K_h\left(x_m - x_r\right) - \theta_0 \sum_{m=0}^{M} K_h\left(x_m - x_r\right) - \theta_1 \sum_{m=0}^{M} \left(x_m - x_r\right) K_h\left(x_m - x_r\right) = 0$$
$$\sum_{m=0}^{M} \widehat{\lambda}_m \left(x_m - x_r\right) K_h\left(x_m - x_r\right) - \theta_0 \sum_{m=0}^{M} \left(x_m - x_r\right) K_h\left(x_m - x_r\right),$$
$$- \theta_1 \sum_{m=0}^{M} \left(x_m - x_r\right)^2 K_h\left(x_m - x_r\right) = 0.$$

Denotando

$$a_j(x_r) = \sum_{m=0}^{M} \left(x_m - x_r\right)^j K_h\left(x_m - x_r\right), \qquad j = 0, 1, 2$$
$$G_j(x_r) = \sum_{m=0}^{M} \widehat{\lambda}_m \left(x_m - x_r\right)^j K_h\left(x_m - x_r\right), \qquad j = 0, 1$$

y sustituyendo estas expresiones en las anteriores se obtiene

$$G_0(x_r) - \theta_0 a_0(x_r) - \theta_1 a_1(x_r) = 0$$
$$G_1(x_r) - \theta_0 a_1(x_r) - \theta_1 a_2(x_r) = 0.$$

Resolviendo este sistema de ecuaciones se tiene

$$\widehat{\theta}_0 = \frac{a_2(x_r) G_0(x_r) - a_1(x_r) G_1(x_r)}{a_2(x_r) a_0(x_r) - a_1^2(x_r)}$$
$$\widehat{\theta}_1 = \frac{a_0(x_r) G_1(x_r) - a_1(x_r) G_0(x_r)}{a_2(x_r) a_0(x_r) - a_1^2(x_r)}.$$

O equivalentemente

$$\widehat{\lambda}_h(x_r) = \widehat{\theta}_0 = \sum_{m=0}^{M} \left(\frac{a_2(x_r) - a_1(x_r)\left(x_m - x_r\right)}{a_2(x_r) a_0(x_r) - a_1^2(x_r)} \right) K_h\left(x_m - x_r\right) \frac{N_m}{\Delta_m} \tag{3.1}$$

y

$$\widehat{\lambda}'_h(x_r) = \widehat{\theta}_1 = \sum_{m=0}^{M} \left(\frac{a_0(x_r)\left(x_m - x_r\right) - a_1(x_r)}{a_2(x_r) a_0(x_r) - a_1^2(x_r)} \right) K_h\left(x_m - x_r\right) \frac{N_m}{\Delta_m}. \tag{3.2}$$

Las varianzas de los estimadores anteriores pueden calcularse fácilmente bajo la hipótesis de PPNH. En particular, la varianza de $\widehat{\lambda}'_h(x_r)$ se obtiene a partir de la expresión (3.2) de la siguiente forma

$$\mathrm{Var}\left[\widehat{\lambda}'_h(x_r)\right] = \sum_{m=0}^{M} \left(\frac{a_0(x_r)\,(x_m - x_r) - a_1(x_r)}{a_2(x_r)a_0(x_r) - a_1^2(x_r)} \right)^2 (K_h\,(x_m - x_r))^2\, \frac{\mathrm{Var}[N_m]}{\Delta_m^2}.$$

Como N_m sigue una distribución de Poisson de parámetro $\lambda_m\Delta_m$, para cada $m = 0, 1, \ldots, M$, se tiene que $\mathrm{Var}[N_m] = \lambda_m\Delta_m$, y utilizando un estimador empírico de λ_m definido previamente por $\widehat{\lambda}_m = N_m/\Delta_m$ con $m = 0, 1, \ldots, M$, se deduce el siguiente estimador de la anterior varianza

$$\widehat{\mathrm{Var}}\left[\widehat{\lambda}'_h(x_r)\right] = \sum_{m=0}^{M} \left(\frac{a_0(x_r)\,(x_m - x_r) - a_1(x_r)}{a_2(x_r)a_0(x_r) - a_1^2(x_r)} \right)^2 (K_h\,(x_m - x_r))^2\, \frac{N_m}{\Delta_m^2}. \quad (3.3)$$

3.4. Inferencia en un espacio de localización y escala

3.4.1. La herramienta exploratoria SiZer Map de Chaudhuri y Marron (1999)

SiZer Map es una herramienta exploratoria gráfica introducida por Chaudhuri y Marron (1999) que permite descubrir características significativas en los datos. SiZer responde a la siguiente cuestión en el análisis de datos: "¿qué características observadas en los datos están verdaderamente ahí?". En el análisis de datos reales, esta cuestión es importante ya que el descubrimiento de características significativas puede dar lugar a importantes conclusiones científicas.

La herramienta SiZer está inspirada en las ideas de la visualización en un espacio de localización y escala desarrolladas en el campo de la computación (ver Lindeberg (1994)). A diferencia de la inferencia no paramétrica tradicional que basa sus conclusiones en estimadores no paramétricos calculados por un nivel óptimo de suavizado, SiZer considera un amplio rango de valores para el parámetro de

suavizado. Se pasa ahí de una inferencia global y en una escala (nivel de suavizado) a una inferencia en un espacio de localización (local) y un espacio de escala (multi-escala).

SiZer fue descrito originalmente para la inferencia visual sobre densidades y funciones de regresión. Para detectar las características significativas, se utilizan múltiples tests basados en intervalos de confianza para las derivadas de la curva subyacente.

A continuación, se describe de forma breve la herramienta SiZer para el problema de la densidad. Se parte de una muestra aleatoria simple $X_1, X_2, \ldots, X_n$, y se considera un estimador tipo núcleo de la función de densidad $f(x)$ de la forma

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

donde h es el parámetro de suavizado que controla la suavidad de la estimación resultante.

Cada pixel del SiZer Map representa un punto indexado por la localización en el eje horizontal y por el parámetro de suavizado en el eje vertical. Los valores del parámetro de suavizado se eligen en función del rango $[h_{\min}, h_{\max}]$. Dicho mapa se basa en intervalos de confianza para $f'_h(x)$ dados por

$$\left(\widehat{f}'_h(x) \pm q\sqrt{\widehat{\mathrm{Var}}[\widehat{f}'_h(x)]}\right), \tag{3.4}$$

donde q es un cuantil apropiado y $\sqrt{\widehat{\mathrm{Var}}[\widehat{f}'_h(x)]}$, es una estimación de la desviación estándar de $\widehat{f}'_h(x)$. Para un nivel de resolución h, el valor del estimador $\widehat{f}'_h(x)$ es significativamente positivo cuando todos los puntos del intervalo (3.4) son positivos y será negativo cuando todos los puntos del intervalo de confianza sean negativos.

SiZer Map muestra información sobre el signo del estimador $\widehat{f}'_h(x)$. Las zonas que se muestran en el SiZer Map son códigos de colores donde el color azul muestra que la derivada es significativamente positiva, el color rojo muestra que la derivada

es significativamente negativa, los puntos en los que la derivada no es significativamente ni positiva ni negativa aparecen de color púrpura y el color gris indica regiones en las que los datos son muy escasos para realizar la estimación.

Para la elección del tipo de cuantiles a utilizar en la estimación de los intervalos de confianza, se incluyen los siguientes cuantiles. Los cuantiles de tipo puntual construidos mediante la aproximación Normal, los de tipo simultáneo construidos mediante la aproximación Normal, los de tipo puntual construidos mediante la aproximación bootstrap y los de tipo simultáneo construidos mediante la aproximación bootstrap. El cálculo de dichos cuantiles aparece detallado en la siguiente sección para la función de intensidad.

3.4.2. SiZer Map para la función de intensidad de un PPNH

En la sección anterior se ha descrito la herramienta SiZer para la densidad propuesta por Chaudhuri y Marron (1999). A continuación se extiende esta herramienta a la función de intensidad, lo que constituye el objetivo principal de este capítulo.

Un ejemplo ilustrativo

A modo ilustrativo, antes de entrar en más detalles, se muestra en la Figura 3.1, el análisis SiZer correspondiente a unos datos que consisten en tiempos entre fallos sucesivos de un sistema hidráulico. En la Sección 3.5.2 se proporcionarán más detalles sobre estos datos. En la gráfica superior se encuentra el Family Plot que muestra las funciones de intensidad estimadas a lo largo del tiempo (localizaciones) con diferentes parámetros de suavizado (escalas). Valores pequeños de los parámetros de suavizado corresponden a grandes resoluciones de visualización que producen estimaciones irregulares de la intensidad con varios falsos cambios de tendencia como resultado de la variabilidad de los datos. En este caso, el conjunto de datos es pequeño con 151 observaciones. Por otro lado, una baja resolución en la

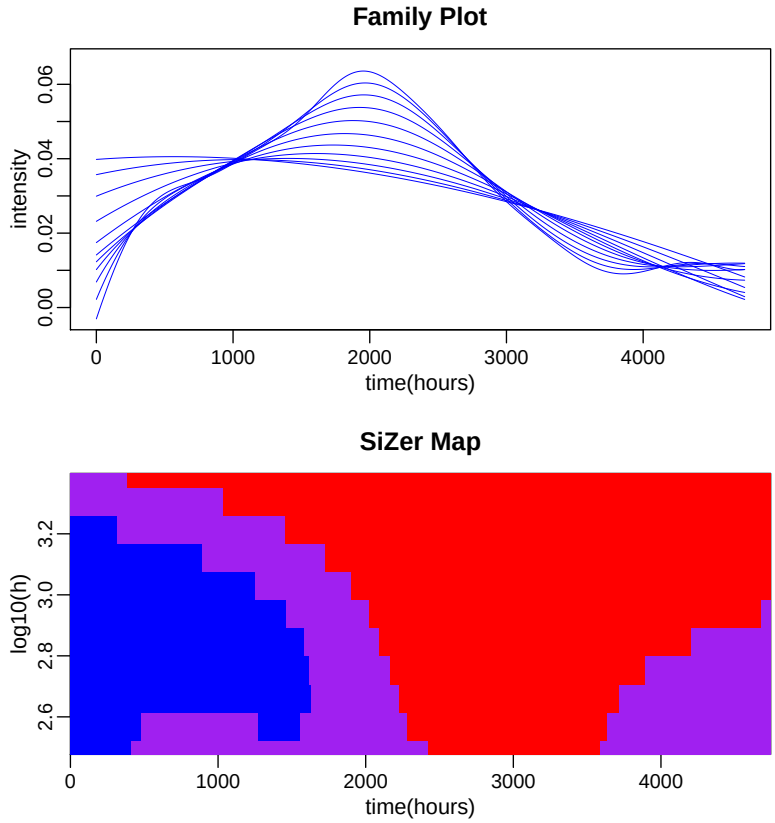


Figura 3.1: Análisis SiZer con intervalos de confianza bootstrap de tipo simultáneo para datos de un sistema hidráulico.

visualización de la función de intensidad subyacente con parámetros de suavizado grandes, tiende a camuflar muchas características que subyacen en la información muestral. La cuestión es si las características mostradas en el Family Plot son en realidad significativas. La gráfica inferior proporciona la respuesta estadística a esta cuestión con un nivel de significación del 5 %. Específicamente, para cada pixel (definido por el tiempo y un parámetro de suavizado) se muestra el signo de la derivada de la función de intensidad con un nivel de confianza del 95 %. Observando el mapa y teniendo en cuenta el código de colores descrito anteriormente, se puede concluir que la intensidad es creciente hasta las 1800 horas y decreciente después.

El cambio de color azul-púrpura-rojo alrededor de las 2000 horas significa que la intensidad realiza un cambio de tendencia (un pico) en ese tiempo. Finalmente, se aprecia que este cambio de tendencia puede visualizarse para la mayoría de los parámetros de suavizado considerados.

Elementos para la construcción del SiZer

Chaudhuri y Marron (1999), mostraron el procedimiento para construir intervalos de confianza para la derivada de la función de densidad. En el caso de la función de intensidad $\lambda(t)$, los intervalos de confianza se pueden definir de forma similar como

$$\left(\widehat{\lambda}'_h(t) + q_{1-\alpha/2}\sqrt{\widehat{\text{Var}}\left[\widehat{\lambda}'_h(t)\right]}, \widehat{\lambda}'_h(t) + q_{\alpha/2}\sqrt{\widehat{\text{Var}}\left[\widehat{\lambda}'_h(t)\right]}\right), \tag{3.5}$$

donde α es el nivel de significación y $q_{1-\alpha/2}$ y $q_{\alpha/2}$ son cuantiles apropiados. En las secciones anteriores se han descrito los principales elementos para poder hacer inferencia relativa a la intensidad de un PPNH con la herramienta SiZer. En concreto, se considera el estimador lineal local de la función de intensidad definido en (3.1), para definir la familia de suavizadores representado en el Family Plot. Los intervalos de confianza para la derivada se definen con el estimador lineal local (3.2), y la estimación de su varianza dada en (3.3). El único aspecto a tratar es el cálculo de los cuantiles.

1. **Cálculo de los cuantiles de tipo puntual q_1 construidos mediante la aproximación Normal**

   Se supone que

   $$\frac{\widehat{\lambda}'_h(t) - \lambda'_h(t)}{\sqrt{\widehat{\text{Var}}\left[\widehat{\lambda}'_h(t)\right]}} \sim \mathcal{N}(0,1),$$

   por lo tanto, dado el nivel de significación α, una aproximación del intervalo de confianza al nivel $(1-\alpha)$ se obtiene tomando $q_{1-\alpha/2}$ y $q_{\alpha/2}$ igual a los cuantiles $(1-\alpha/2)$ y $\alpha/2$ de la distribución Normal estándar, respectivamente.

2. **Cálculo de los cuantiles de tipo simultáneo q_2 construidos mediante la aproximación Normal**

   El problema de comparación múltiple que subyace en SiZer se puede tratar considerando intervalos de confianza de tipo simultáneo. Chaudhuri y Marron (1999) proponen los cuantiles

   $$q_{1-\alpha/2} = -q_{\alpha/2} = \Phi^{-1}\left(\frac{1+(1-\alpha)^{1/m(h)}}{2}\right),$$

   donde Φ^{-1} es la inversa de la función de distribución Normal estándar, α es el nivel de significación elegido y $m(h)$ es el número de bloques independientes de tamaño promedio disponibles de un conjunto de datos de tamaño $N(\tau)$. Concretamente, $m(h)$ se define como

   $$m(h) = \frac{N(\tau)}{\text{avg}_{t\in\mathcal{D}_h}\text{ESS}(t,h)},$$

   donde $\text{ESS}(t,h)$ es el tamaño efectivo de la muestra que viene dado por

   $$\text{ESS}(t,h) = \frac{\sum_{i=1}^{N(\tau)} K_h(t-T_i)}{K_h(0)},$$

   y $\text{avg}_{t\in\mathcal{D}_h}\text{ESS}(t,h)$ indica el valor promedio de $\text{ESS}(t,h)$ sobre el conjunto de localizaciones en el que los datos son densos $\mathcal{D}_h = \{t : \text{ESS}(t,h) \geq 5\}$. El $\text{ESS}(t,h)$ se utiliza para indicar al suavizado dónde hay escasez de datos para poder realizar la estimación efectiva.

3. **Cálculo de los cuantiles de tipo puntual q_3 construidos mediante la aproximación bootstrap**

   A través de la aproximación bootstrap, pueden obtenerse intervalos de confianza bastante precisos sin asumir normalidad, llamados "bootstrap-t" (ver Efron y Tibshirani (1993), Capítulo 12). El siguiente algoritmo resume el cálculo de los intervalos de confianza de tipo puntual construidos mediante la aproximación bootstrap:

Paso 1 Se generan B muestras bootstrap.

Paso 2 Para $b = 1, 2, \ldots, B$ se calcula

$$Z_b^*(t,h) = \frac{\widehat{\lambda}_h'(t)^{*b} - \widehat{\lambda}_h'(t)}{\sqrt{\widehat{\text{Var}}\left[\widehat{\lambda}_h'(t)^{*b}\right]}}, \quad \forall t \in [0,\tau]$$

donde $\widehat{\lambda}_h'(t)^{*b}$ es el valor de $\widehat{\lambda}_h'(t)$ para las b muestras bootstrap y $\sqrt{\widehat{\text{Var}}\left[\widehat{\lambda}_h'(t)^{*b}\right]}$ es un estimador de la desviación estándar de $\widehat{\lambda}_h'(t)^{*b}$.

Paso 3 Los cuantiles $q_{1-\alpha/2}$ y $q_{\alpha/2}$ vienen dados por los cuantiles $(1-\alpha/2)$ y $\alpha/2$ de la muestra $Z_1^*(t,h), Z_2^*(t,h), \ldots, Z_B^*(t,h)$, respectivamente.

Para la generación de las muestras bootstrap se procede según Cowling *et al.* (1996). Se generan n^* tiempos condicionados a los datos a partir de una distribución de Poisson con parámetro $\widehat{\Lambda}(\tau) = \int_0^\tau \widehat{\lambda}(u)du$ y se obtienen los tiempos $T_1^*, T_2^*, \ldots, T_{n^*}^*$ mediante un muestreo aleatorio con reemplazamiento de tamaño n^* a partir de los datos $T_1, T_2, \ldots, T_n$. Este procedimiento viene motivado por el hecho de que, condicionado al evento $N(\tau) = n$, los $T_1, T_2, \ldots, T_n$ son los tiempos ordenados de una muestra de tamaño n obtenidos de una distribución con densidad $f(t) = \lambda(t)/\Lambda(\tau)$. El estimador bootstrap $\widehat{\lambda}_h(t)^*$ de $\widehat{\lambda}(t)$ viene definido en (3.1) y el estimador de su derivada $\widehat{\lambda}_h'(t)^*$ en (3.2).

4. **Cálculo de los cuantiles de tipo simultáneo q_4 construidos mediante la aproximación bootstrap**

Los intervalos de confianza de tipo simultáneo construidos mediante la aproximación bootstrap se calculan mediante el siguiente algoritmo:

Paso 1 Se generan B muestras bootstrap.

Paso 2 Para cada $b = 1, 2, \ldots, B$ se calcula

$$Z_{\text{inf}}^{*b} = \inf_{t \in \mathcal{D}_h} \{Z_b^*(t,h)\}$$

$$Z^{*b}_{\text{sup}} = \sup_{t \in \mathcal{D}_h} \{Z^*_b(t,h)\}$$

Paso 3 Los cuantiles $q_{1-\alpha/2}$ vienen dados por los cuantiles $(1-\alpha/2)$ de la muestra $Z^{*1}_{\text{sup}}, Z^{*2}_{\text{sup}}, \ldots, Z^{*B}_{\text{sup}}$ y los $q_{\alpha/2}$ vienen dados por los cuantiles $\alpha/2$ de la muestra $Z^{*1}_{\text{inf}}, Z^{*2}_{\text{inf}}, \ldots, Z^{*B}_{\text{inf}}$.

donde $\mathcal{D}_h$ y las muestras bootstrap se generan de forma análoga al caso puntual descrito antes.

3.5. Estudio de Simulación

En esta sección se describe un estudio de simulación cuya finalidad es la de evaluar la herramienta gráfica SiZer Map descrita anteriormente. El comportamiento de SiZer se evalúa atendiendo a los dos aspectos siguientes: (1) comprobar en qué medida SiZer Map permite detectar las características que realmente subyacen en los datos; y (2) evaluar la cobertura de los intervalos de confianza utilizados por SiZer para extraer conclusiones acerca de la estructura de los datos.

3.5.1. Descripción de los modelos simulados

Se consideran cuatro modelos que permiten ilustrar situaciones donde la función de intensidad presenta estructuras diferentes. La expresión de la función de intensidad en cada modelo así como el tamaño muestral considerado se describen en la Tabla 3.1. En la Figura 3.2 pueden verse las representaciones gráficas de los cuatro modelos. Para cada caso se han generado 1000 muestras aleatorias.

Tabla 3.1: Descripción de los modelos teóricos.

Modelo	Función de intensidad teórica	Tamaño muestral
M1	$\lambda(t) = 1$	100
M2	$\lambda(t) = 3t^2$	500
M3	$\lambda(t) = 100 + 50\cos\left(\frac{1}{2}\pi t\right)$	500
M4	$\lambda(t) = 100 + 50\cos\left(\frac{3}{2}\pi t\right)$	1000

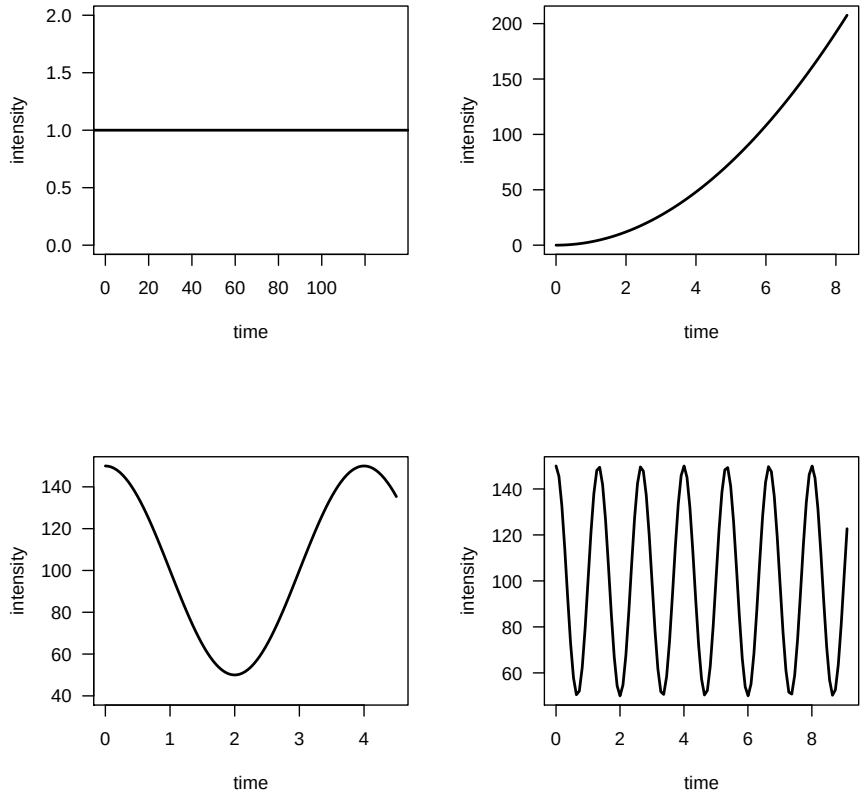


Figura 3.2: Representación gráfica de los cuatro modelos simulados: De izquierda a derecha, la primera fila muestra los modelos M1 y M2, la segunda fila muestra los modelos M3 y M4.

La generación de las muestras se ha hecho como sigue. Para cada modelo se considera $\Lambda(t) = \int_0^t \lambda(u)du$ la función de intensidad acumulada. Los tiempos t_i se generan de la siguiente forma: para $i = 1, 2, \ldots, n$ se genera u_i desde una $\mathcal{U}(0, 1)$, seguidamente se calcula $x_i = -\log u_i$ y se define $r_i = r_{i-1} + x_i$. Aquí se ha considerado $r_0 = 0$. Finalmente los tiempos se definen calculando $t_i = \Lambda^{-1}(r_i)$.

Para la construcción de SiZer se han considerado dos rejillas equiespaciadas, una de localizaciones (tiempos donde se realiza la estimación) y otra de niveles de escala (valores del parámetro de suavizado). La rejilla de localizaciones tiene

tamaño 401 y sus valores mínimo y máximo han sido elegidos para cada modelo de modo que contengan todos los tiempos generados en las muestras consideradas. Respecto a la rejilla de parámetros de suavizado se ha definido en escala logarítmica y considerando 11 valores equiespaciados en un intervalo adecuado, atendiendo de nuevo al rango de los tiempos generados según cada modelo. Los 11 valores considerados en cada caso permiten ofrecer suficientes niveles de escala para la visualización del problema. Como estimador de la derivada de la función de intensidad se ha utilizado el estimador lineal local con el núcleo de Epanechnikov. Por otro lado, los intervalos de confianza para la primera derivada de la función de intensidad que utiliza SiZer para la inferencia han sido calculados utilizando tanto la aproximación Normal como la aproximación bootstrap. En este último caso el número de muestras bootstrap simuladas ha sido de 500. Para ambos tipos de aproximación se han considerado intervalos de tipo puntual y simultáneo.

Una primera intuición acerca del comportamiento de SiZer Map en los cuatro modelos que se han simulado, se puede obtener a partir de los gráficos generados a partir de una de las 1000 muestras seleccionadas para cada caso. El resultado puede verse en las Figuras 3.3, 3.4, 3.5 y 3.6, correspondientes a los modelos M1, M2, M3 y M4, respectivamente.

La Figura 3.3 muestra el Family Plot (primera gráfica) y el SiZer Map (segunda gráfica) correspondientes a una muestra generada desde el modelo M1. En este caso SiZer se ha definido considerando intervalos de tipo bootstrap y simultáneos. El Family Plot muestra las estimaciones de la función de intensidad para cada parámetro de suavizado considerado, donde se han utilizado 11 parámetros de suavizado y por tanto en el gráfico se representan 11 estimaciones de la función de intensidad subyacente. SiZer Map muestra las características significativas que subyacen en los datos simulados a través de un lenguaje de colores. Como se describió anteriormente, el color púrpura representa el caso en que el intervalo de confianza para la derivada de la función de intensidad contiene al cero, el azul el

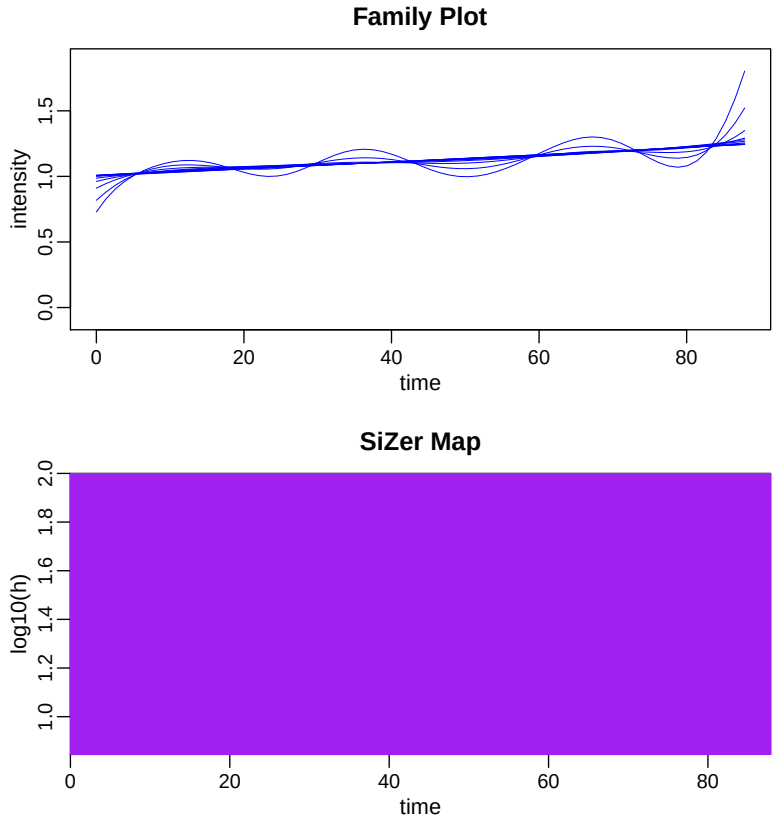


Figura 3.3: Análisis SiZer con intervalos de confianza bootstrap de tipo simultáneo para datos simulados del modelo M1.

caso en que el intervalo es estrictamente positivo y el color rojo el caso en el que el intervalo es estrictamente negativo. El modelo M1 se define con una función de intensidad constante donde la primera derivada es siempre cero. Se puede apreciar por tanto que SiZer detecta perfectamente este caso mostrando un mapa donde el color púrpura indica que no hay ningún cambio significativo en el signo de la derivada para todo el intervalo de localizaciones y de parámetros de suavizado.

La Figura 3.4 corresponde al modelo M2 considerando ahora SiZer con intervalos de tipo puntual construidos mediante la aproximación bootstrap. En este modelo la función de intensidad simulada es estrictamente creciente y por tanto

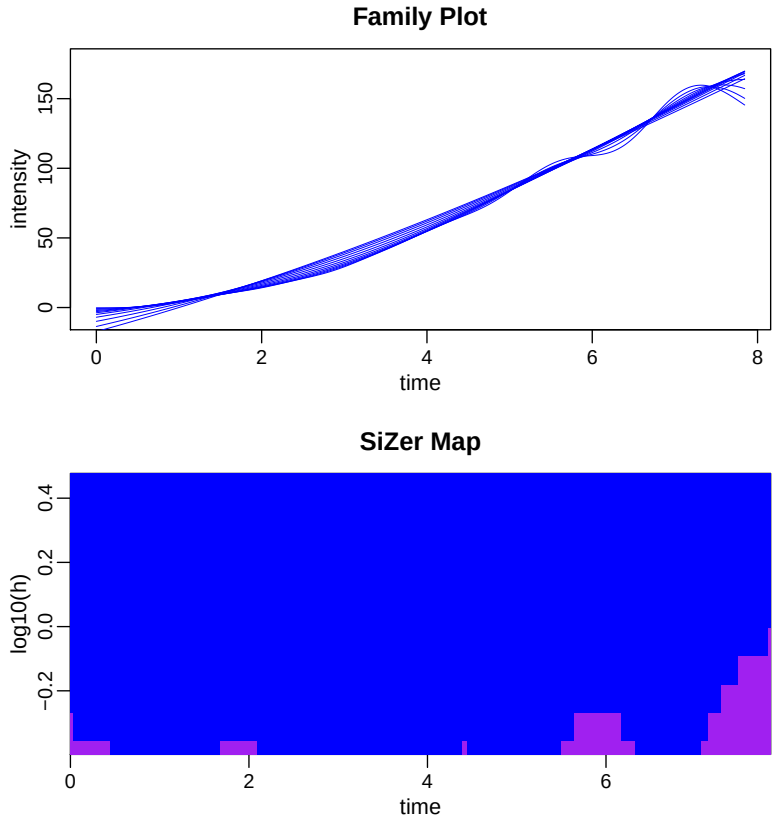


Figura 3.4: Análisis SiZer con intervalos de confianza bootstrap de tipo puntual para datos simulados del modelo M2.

su derivada es estrictamente positiva en todos los tiempos. Esta situación es la que describe en general SiZer Map calculado a partir de la muestra simulada del modelo. De hecho se observa que predomina el color azul en prácticamente todo el mapa. La única excepción está en las localizaciones mayores y para escalas grandes donde SiZer indica con un cambio de color azul-púrpura-rojo que hay una característica significativa en la curva definida por un cambio de creciente a decreciente. Este hecho es debido al efecto de la frontera en la estimación núcleo de la curva.

La Figura 3.5 muestra los resultados para una muestra generada a partir del modelo M3, considerando SiZer con intervalos de tipo puntual construidos me-

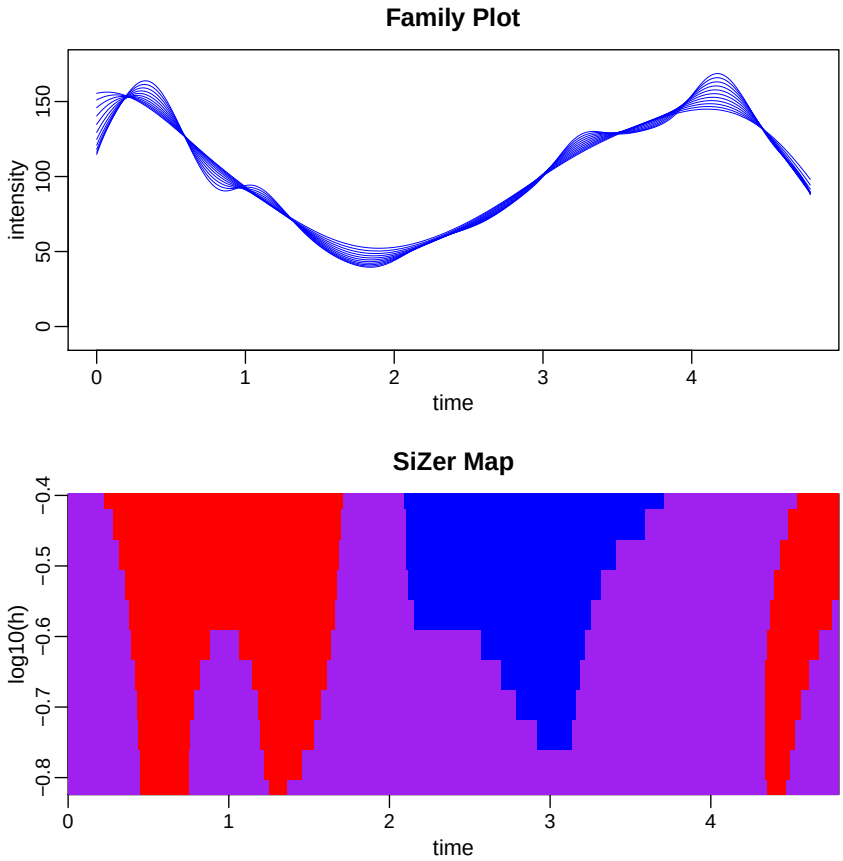


Figura 3.5: Análisis SiZer con intervalos de confianza bootstrap de tipo puntual para datos simulados del modelo M3.

diante la aproximación bootstrap. En este caso, el mapa muestra claramente las dos características significativas que existen en la curva simulada, para todos los niveles de escala mostrados. Tales cambios vienen reflejados por el cambio de color azul-púrpura-rojo y viceversa, a lo largo del espacio de localizaciones.

Finalmente, la Figura 3.6 muestra los resultados para la muestra generada a partir del modelo M4, y considerando SiZer con intervalos de tipo puntual construidos mediante la aproximación Normal. En esta ocasión, el gráfico SiZer Map muestra de forma acertada los trece cambios significativos que la intensidad teórica presenta a lo largo del rango de localizaciones. Estos cambios son detectados para

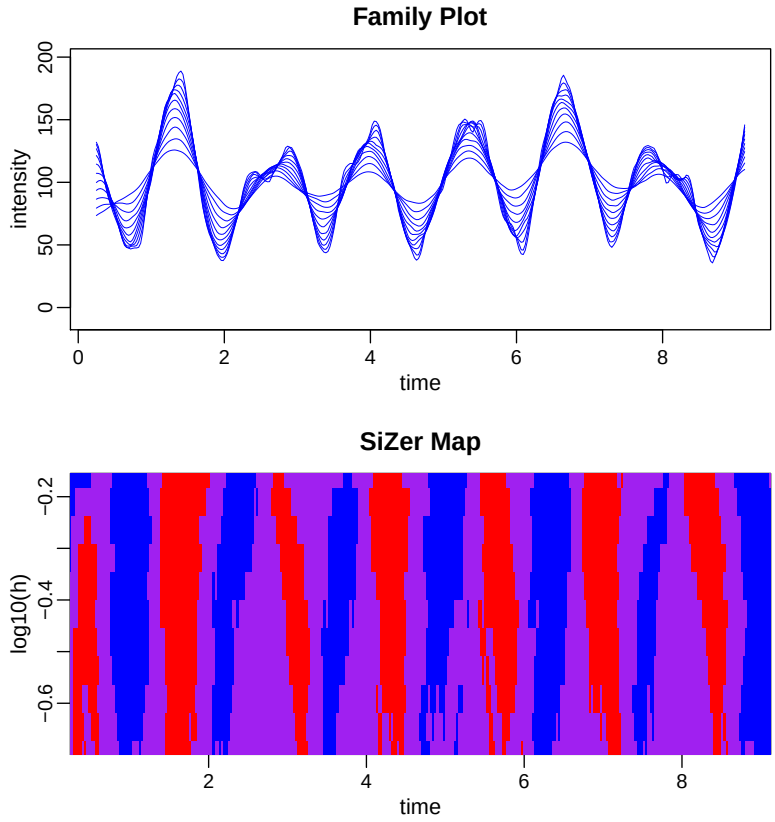


Figura 3.6: Análisis SiZer con intervalos de confianza Normal de tipo puntual para datos simulados del modelo M4.

todos los niveles de escala considerados.

Los buenos resultados que se han descrito para una muestra generada al azar en cada modelo han de confirmarse no obstante eliminando el efecto muestral. A continuación se describe el procedimiento que se ha seguido para evaluar el comportamiento de la herramienta SiZer eliminando el efecto muestral. Como se indicó al comienzo de esta sección se persiguen dos objetivos. En primer lugar, comprobar si SiZer es capaz de detectar las características que realmente tienen las funciones de intensidad simuladas. Para ello se ha calculado el número de cambios de signo significativos en la primera derivada detectados por el SiZer Map para cada valor

del parámetro de suavizado. En segundo lugar, evaluar la cobertura de los intervalos de confianza para la primera derivada de la función de intensidad. Para ello, considerando las 1000 muestras simuladas de cada modelo, se define la cobertura empírica del intervalo en cada localización como el número de veces en que éste contiene al valor de la derivada de la función de intensidad teórica. Los resultados obtenidos se representan gráficamente mediante un mapa de color. En este caso se ha utilizado un rango de colores que oscilan entre el verde (menor cobertura) y el gris claro (mayor cobertura). En este mapa las coberturas se muestran según los parámetros de suavizado considerados en la rejilla en escala logarítmica (eje de ordenadas) y los tiempos en la rejilla de estimación (eje de abcisas).

3.5.2. Evaluación de SiZer para la detección de cambios en la tendencia

A continuación se muestran los resultados de simulación relativos al primero de los objetivos. Los resultados para el modelo M1 se recogen en la Tabla 3.2. Dicha tabla muestra el número de cambios de signos significativos en la primera derivada detectados por el SiZer Map, construido mediante la aproximación Normal y bootstrap, utilizando intervalos de tipo puntual así como simultáneo. Los recuentos se muestran para cada uno de los valores del parámetro de suavizado considerados. Se puede apreciar que los intervalos de tipo simultáneo en SiZer Map ofrecen mejores resultados que los de tipo puntual, cuando el objetivo es detectar los cambios significativos en la función de intensidad. De hecho para casi todas las muestras generadas y los parámetros de suavizado considerados, SiZer con intervalos simultáneos es capaz de detectar el número correcto de cambios de signo en la derivada, que en este modelo es cero. Además, considerando intervalos de confianza simultáneos, se puede ver que SiZer Map construido a partir de la metodología bootstrap ofrece resultados ligeramente mejores que en el caso de usar la aproximación Normal.

Conclusiones similares se pueden extraer analizando los resultados para el modelo M2 recogidos en la Tabla 3.3. De nuevo SiZer Map con intervalos simultáneos, y especialmente los basados en la metodología bootstrap, son capaces de detectar en gran medida el número correcto de cambios de signo significativos en la primera derivada, que en este modelo es de nuevo cero al ser la función de intensidad simulada estrictamente creciente.

Por otro lado, los resultados obtenidos para los modelos M3 y M4 conducen a conclusiones diferentes como puede verse en las Tablas 3.4 y 3.5. En este caso se trata de modelos de más complejidad, en concreto el modelo M4 es un modelo de alta frecuencia. SiZer Map considerando intervalos de tipo puntual ofrece mejores resultados que considerando intervalos de tipo simultáneo. Esto ocurre tanto con la aproximación bootstrap como en el caso de considerar la aproximación Normal. Los intervalos de tipo simultáneo tienden a ocultar las características de la función subyacente cuando estas se presentan en localizaciones cercanas. Este hecho es claramente evidente en el caso del modelo M4 que presenta trece cambios de signo en el rango de estimación considerado, no obstante también se puede apreciar en el modelo M3 que presenta tres cambios de signo. En las Tablas 3.2 y 3.5 aparecen un pequeño número de muestras descartadas, esto es, muestras en las que no se han podido construir los intervalos de confianza debido a que los datos están muy dispersos para ese nivel de suavizado.

3.5.3. Cobertura de los intervalos de confianza

A continuación, se analizan los resultados obtenidos relativos a la cobertura de los intervalos de confianza utilizados por el análisis SiZer. La Figura 3.7 muestra los porcentajes de cobertura empírica para los intervalos de confianza de tipo puntual construidos mediante las aproximaciones Normales y bootstrap según el parámetro de suavizado considerado y las localizaciones, para el modelo M1. La Figura 3.7, muestra mejores resultados para los intervalos de confianza de tipo puntual

Tabla 3.2: Recuentos observados en 1000 muestras de tamaño $n = 100$ simuladas a partir del modelo M1. Se muestran por columnas los niveles de suavizado considerados en escala logarítmica. Las filas en color gris muestran el número de veces que SiZer detecta el número real de cambios en la tendencia de la función de intensidad.

Modelo M1	$\log_{10}$(h) / Núm. cambios	1.00	1.05	1.10	1.14	1.19	1.24	1.29	1.33	1.38	1.43	1.48
Puntual Normal	0	687	741	769	790	804	835	860	869	822	794	742
	1	225	205	196	172	167	147	129	122	171	202	250
	2	58	42	33	37	29	18	11	9	7	4	8
	3	9	8	2	1	0	0	0	0	0	0	0
	4	1	0	0	0	0	0	0	0	0	0	0
Simultáneo Normal	0	967	977	979	979	980	975	973	969	944	926	870
	1	10	18	20	19	19	24	27	31	56	74	130
	2	3	1	1	2	1	1	0	0	0	0	0
Puntual bootstrap	0	452	532	589	630	670	724	762	783	743	705	658
	1	340	320	314	295	272	239	217	201	242	283	326
	2	146	120	88	73	57	37	21	16	15	12	16
	3	38	23	9	2	1	0	0	0	0	0	0
	4	4	1	1	0	0	0	0	0	0	0	0
Simultáneo bootstrap	0	980	996	1000	1000	1000	1000	1000	1000	1000	999	992
	1	0	0	0	0	0	0	0	0	0	1	8
	Muestras descartadas	20	4	0	0	0	0	0	0	0	0	0

construidos mediante la aproximación Normal que en los construidos mediante la aproximación bootstrap. No obstante, en estos casos, el verdadero interés se centra en analizar las coberturas de los intervalos de tipo simultáneo, que como se dijo anteriormente, se comportan mejor en los modelos M1 y M2.

Además de la cobertura empírica para cada localización, se ha evaluado la cobertura empírica total, definida como el número de veces en que todos los intervalos de la derivada de la función de intensidad teórica en cada localización contienen a los verdaderos valores. En este caso las coberturas empíricas son valores globales obtenidos para cada parámetro de suavizado considerado. La Figura 3.8 muestra los porcentajes de cobertura empírica total observada en el modelo M1. El porcentaje de cobertura se describe a través de una curva y en estos gráficos se han comparado las curvas correspondientes a cada una de las cuatro formas en las que se construye SiZer. De este gráfico se puede apreciar que la cobertura es mayor para los intervalos de tipo simultáneo y en especial cuando se considera la metodología bootstrap.

Tabla 3.3: Recuentos observados en 1000 muestras de tamaño $n = 500$ simuladas a partir del modelo M2. Se muestran por columnas los niveles de suavizado considerados en escala logarítmica. Las filas en color gris muestran el número de veces que SiZer detecta el número real de cambios en la tendencia de la función de intensidad.

Modelo M2	$\log_{10}$(h) / Núm. cambios	-0.10	-0.06	-0.02	0.02	0.06	0.10	0.14	0.18	0.22	0.26	0.30
Puntual Normal	0	998	1000	1000	1000	1000	1000	1000	995	980	936	820
	1	0	0	0	0	0	0	0	5	20	64	180
	2	2	0	0	0	0	0	0	0	0	0	0
Simultáneo Normal	0	1000	1000	1000	1000	1000	1000	1000	996	986	963	873
	1	0	0	0	0	0	0	0	4	14	37	127
Puntual bootstrap	0	993	997	999	1000	1000	1000	997	994	973	914	785
	1	2	1	0	0	0	0	3	6	27	86	215
	2	5	2	1	0	0	0	0	0	0	0	0
Simultáneo bootstrap	0	1000	1000	1000	1000	1000	1000	1000	1000	1000	997	995
	1	0	0	0	0	0	0	0	0	0	3	5

Para el modelo M2 se obtienen conclusiones similares al modelo M1. En la Figura 3.9, se observa que la franja de gris claro es mayor en los intervalos de confianza de tipo puntual construidos mediante la aproximación Normal que en la aproximación bootstrap. No obstante, como sucedía en el modelo M1, el verdadero interés se centra en analizar las coberturas de los intervalos de confianza de tipo simultáneo. La Figura 3.10 muestra que la cobertura empírica es mayor para los casos de intervalos de confianza de tipo simultáneo, especialmente para los intervalos de confianza construidos mediante la aproximación bootstrap.

En el caso de los modelos M3 y M4 se muestran únicamente las gráficas de las coberturas a partir de intervalos de tipo puntual, ya que como se vio anteriormente en las Tablas 3.4 y 3.5, el comportamiento de SiZer con intervalos de tipo puntual mejora notablemente con respecto a los simultáneos. En la Figura 3.11 se muestran las gráficas de las coberturas empíricas para cada localización y escala, tanto para la aproximación Normal como para la aproximación bootstrap. La aproximación Normal ofrece resultados ligeramente mejores que la aproximación bootstrap en este caso. Algo similar ocurre con el modelo M4 representado en la Figura 3.12.

Tabla 3.4: Recuentos observados en 1000 muestras de tamaño $n = 500$ simuladas a partir del modelo M3. Se muestran por columnas los niveles de suavizado considerados en escala logarítmica. Las filas en color gris muestran el número de veces que SiZer detecta el número real de cambios en la tendencia de la función de intensidad.

Modelo M3	$\log_{10}$(h) / Núm. cambios	-0.30	-0.26	-0.22	-0.18	-0.14	-0.09	-0.05	-0.01	0.03	0.07	0.11
Puntual Normal	1	30	17	13	3	2	0	0	0	0	0	1
	2	966	983	987	997	998	999	994	975	851	555	240
	3	0	0	0	0	0	1	6	25	149	445	759
	4	4	0	0	0	0	0	0	0	0	0	0
Simultáneo Normal	0	58	28	9	0	0	0	0	0	0	0	0
	1	144	79	37	18	5	1	0	0	0	0	2
	2	798	893	954	982	995	999	999	993	958	751	434
	3	0	0	0	0	0	0	1	7	42	249	564
Puntual bootstrap	1	24	15	9	3	1	0	0	0	0	0	0
	2	970	983	991	997	999	999	994	971	797	473	174
	3	1	0	0	0	0	1	6	29	203	527	826
	4	5	2	0	0	0	0	0	0	0	0	0
Simultáneo bootstrap	0	665	521	369	217	131	59	24	11	1	0	0
	1	184	200	177	124	94	47	27	13	9	14	24
	2	151	279	454	659	775	894	949	975	989	975	921
	3	0	0	0	0	0	0	0	1	1	11	55

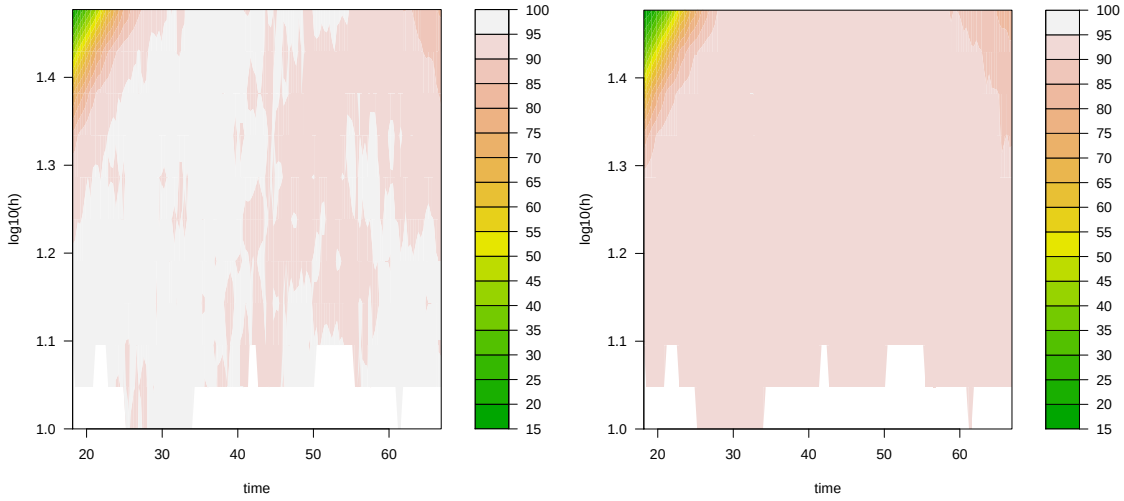
3.5.4. Resultados y conclusiones

El objetivo principal de la herramienta SiZer es la de detectar las características que realmente están presentes en la función de intensidad que subyace en los datos. En este sentido y viendo los resultados derivados de las simulaciones, parece más adecuado el uso de la aproximación bootstrap que la aproximación Normal. Sin embargo, un inconveniente importante a considerar de la aproximación bootstrap es su alto coste computacional, por lo que sólo se recomienda en caso donde la mejora sea sustancial o donde el tamaño reducido de la muestra de datos no justifique la aproximación Normal.

Por otro lado, se concluye que el uso de la metodología simultánea es más apropiada para modelos sencillos sin muchas características en la función de intensidad como son los modelos M1 y M2. En cambio, según el estudio de simulación, se muestra que para modelos más complejos con muchas características como son los modelos M3 y M4 sería más apropiada la metodología puntual. Estas conclusiones

Tabla 3.5: Recuentos observados en 1000 muestras de tamaño $n = 1000$ simuladas a partir del modelo M4. Se muestran por columnas los niveles de suavizado considerados en escala logarítmica. Las filas en color gris muestran el número de veces que SiZer detecta el número real de cambios en la tendencia de la función de intensidad.

Modelo M4	$\log_{10}$(h) / Núm. cambios	-0.70	-0.64	-0.59	-0.54	-0.48	-0.43	-0.37	-0.32	-0.26	-0.21	-0.15
Puntual Normal	5	1	0	0	0	0	0	0	0	0	0	0
	6	5	0	0	0	0	0	0	0	0	0	6
	7	32	8	5	0	1	0	0	0	0	1	3
	8	36	17	2	0	0	0	0	0	0	4	80
	9	140	68	37	17	9	4	2	1	1	6	71
	10	88	43	25	22	7	5	5	9	12	92	321
	11	306	312	230	177	142	69	43	24	27	59	88
	12	82	85	81	73	73	62	85	201	459	747	426
	13	301	465	616	707	760	838	768	571	264	29	3
	14	0	0	3	3	8	22	97	192	234	60	1
	15	3	1	1	1	0	0	0	2	3	2	0
Simultáneo Normal	0	427	148	34	9	1	0	0	0	0	0	0
	1	309	263	106	28	10	0	0	0	0	0	0
	2	162	249	184	75	21	0	0	0	0	0	5
	3	70	181	217	154	54	3	1	0	0	0	17
	4	21	102	183	155	91	12	2	1	1	3	68
	5	3	36	143	180	152	30	5	2	1	9	86
	6	2	15	70	132	119	57	15	5	10	48	212
	7	0	4	44	125	207	137	51	20	19	43	142
	8	0	1	13	71	120	119	72	59	82	159	256
	9	0	0	4	48	129	212	126	75	55	79	85
	10	0	0	2	15	50	121	161	208	287	381	102
	11	0	0	0	6	36	197	250	198	109	59	11
	12	0	0	0	2	4	47	138	268	366	218	16
	13	0	0	0	0	6	65	179	163	68	1	0
	14	0	0	0	0	0	0	0	1	2	0	0
Puntual bootstrap	5	0	0	0	0	0	0	0	0	0	0	1
	6	0	0	0	0	0	0	0	0	0	0	4
	7	2	0	0	0	0	0	0	0	0	0	6
	8	4	1	0	0	0	0	0	0	0	3	61
	9	31	11	7	3	3	1	1	0	0	7	65
	10	18	15	7	4	4	5	3	3	9	62	286
	11	218	162	136	95	63	45	20	12	16	49	89
	12	77	64	55	47	51	45	71	178	415	763	482
	13	624	738	790	846	870	865	752	526	246	30	2
	14	5	3	3	4	9	39	153	277	309	83	4
	15	15	5	2	1	0	0	0	4	5	3	0
Simultáneo bootstrap	0	972	882	694	453	250	53	6	2	2	4	54
	1	21	93	225	300	276	87	15	8	4	18	126
	2	1	24	58	158	208	163	53	32	16	51	200
	3	0	0	20	59	146	186	96	37	52	90	212
	4	0	0	3	27	70	169	137	87	89	152	191
	5	0	0	0	2	36	140	156	103	94	160	99
	6	0	0	0	1	5	84	177	173	186	180	77
	7	0	0	0	0	6	69	139	148	138	128	25
	8	0	0	0	0	2	19	92	175	194	134	15
	9	0	0	0	0	1	24	77	102	89	37	1
	10	0	0	0	0	0	5	29	77	88	36	0
	11	0	0	0	0	0	1	18	33	23	4	0
	12	0	0	0	0	0	0	4	17	23	6	0
	13	0	0	0	0	0	0	1	6	2	0	0
	Muestras descartadas	6	1	0	0	0	0	0	0	0	0	0



(a) (b)

Figura 3.7: Cobertura empírica puntual. a) Cobertura de los Intervalos de confianza de tipo puntual construidos mediante la aproximación Normal. b) Cobertura de los Intervalos de confianza de tipo puntual construidos mediante la aproximación bootstrap. Resultados con 1000 muestras de tamaño $n = 100$ del modelo M1.

fueron también derivadas del estudio de Oliveira (2013) para el caso de densidades y funciones de regresión con datos circulares.

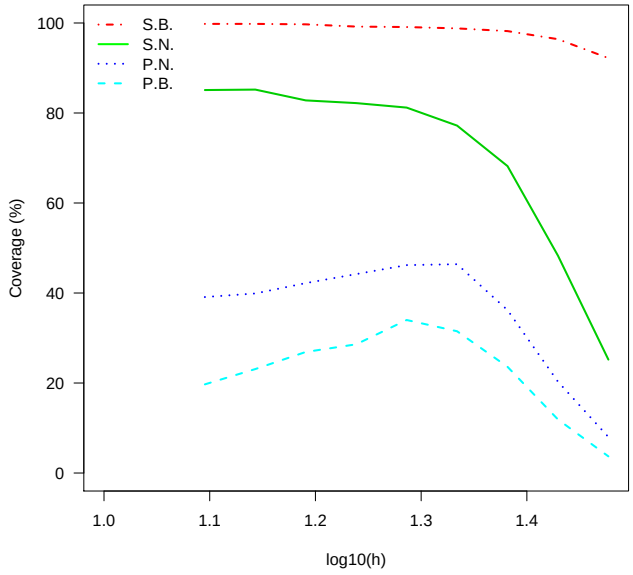
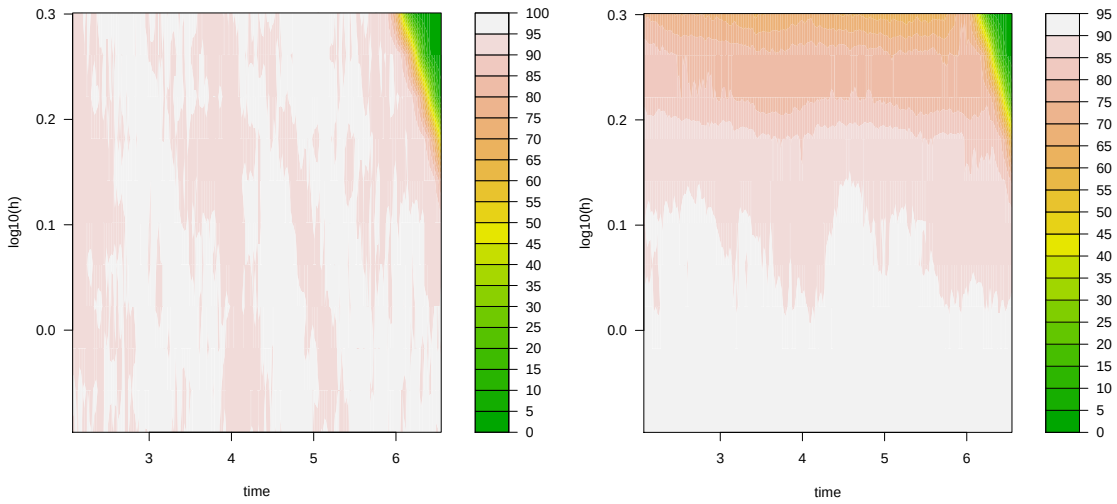


Figura 3.8: Cobertura empírica global. S.B. denota intervalos bootstrap de tipo simultáneo. S.N. intervalos de tipo simultáneo construidos mediante la aproximación Normal. P.N. intervalos de tipo puntual construidos mediante la aproximación Normal. P.B. intervalos bootstrap de tipo puntual. Muestras de tamaño $n = 100$ del modelo M1.



(a) (b)

Figura 3.9: Cobertura empírica puntual. a) Cobertura de los Intervalos de confianza de tipo puntual construidos mediante la aproximación Normal. b) Cobertura de los Intervalos de confianza de tipo puntual construidos mediante la aproximación bootstrap. Resultados con 1000 muestras de tamaño $n = 500$ del modelo M2.

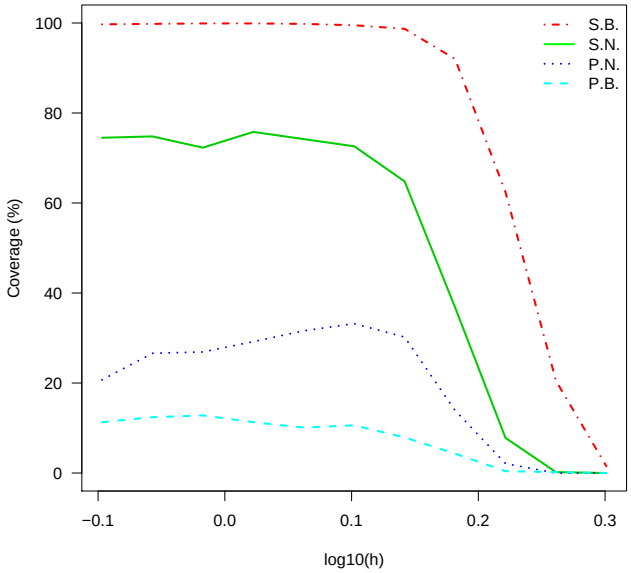
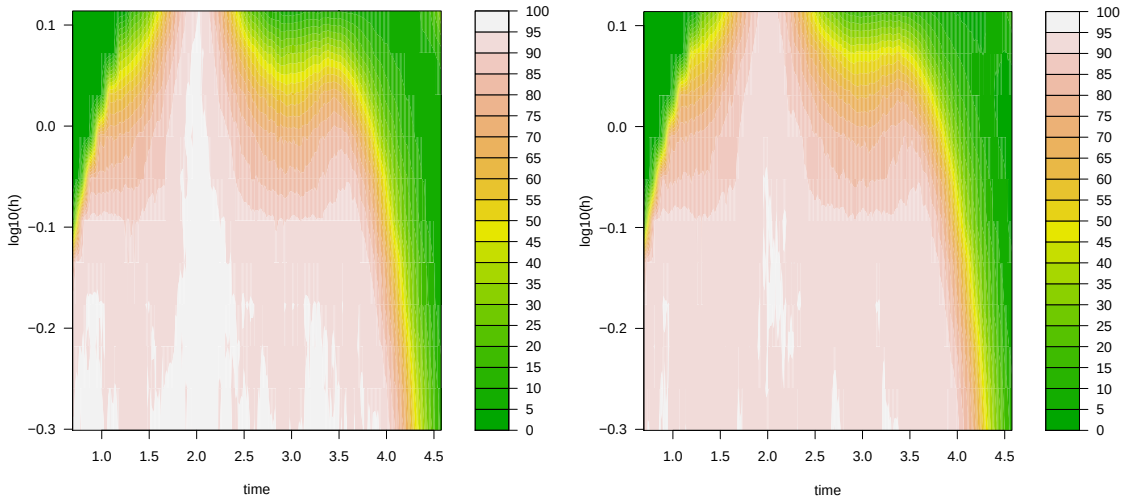
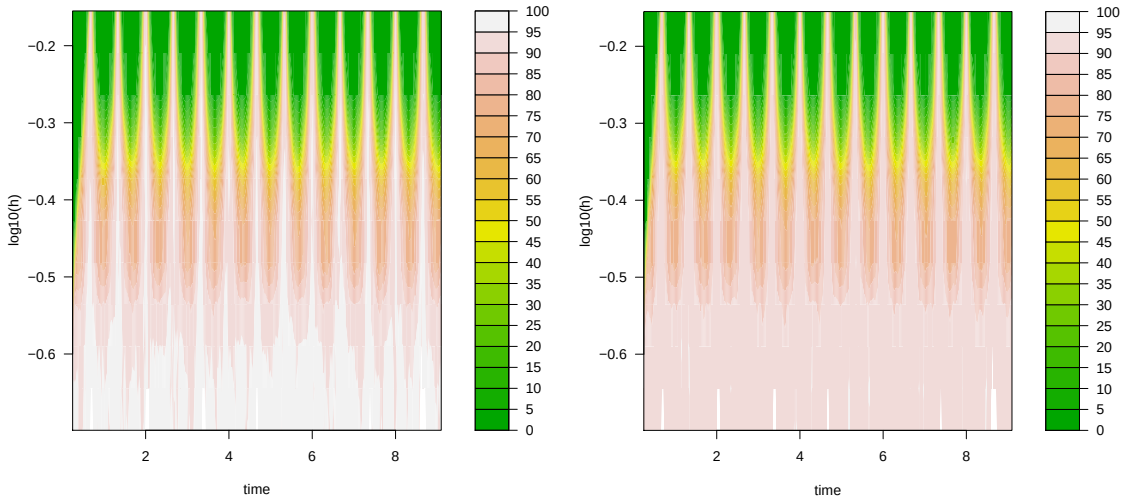


Figura 3.10: Cobertura empírica global. S.B. denota intervalos bootstrap de tipo simultáneo. S.N. intervalos de tipo simultáneo construidos mediante la aproximación Normal. P.N. intervalos de tipo puntual construidos mediante la aproximación Normal. P.B. intervalos bootstrap de tipo puntual. Muestras de tamaño $n = 500$ del modelo M2.



(a) (b)

Figura 3.11: Cobertura empírica puntual. a) Cobertura de los Intervalos de confianza de tipo puntual construidos mediante la aproximación Normal. b) Cobertura de los Intervalos de confianza de tipo puntual construidos mediante la aproximación bootstrap. Resultados con 1000 muestras de tamaño $n = 500$ del modelo M3.



(a) (b)

Figura 3.12: Cobertura empírica puntual. a) Cobertura de los Intervalos de confianza de tipo puntual construidos mediante la aproximación Normal. b) Cobertura de los Intervalos de confianza de tipo puntual construidos mediante la aproximación bootstrap. Resultados con 1000 muestras de tamaño $n = 1000$ del modelo M4.

3.6. Aplicaciones con datos reales

3.6.1. Datos de ocurrencia de réplicas en terremotos

La primera aplicación a datos reales consiste en el análisis de unos datos relativos a réplicas de terremotos que suceden después de un terremoto de gran magnitud. En los casos que se han analizado, la ocurrencia de réplicas puede considerarse como un proceso puntual en el tiempo, concretamente, un PPNH. El principal interés se centra en analizar las características más importantes de la intensidad subyacente del proceso. En particular, la ley de Omori modificada, $\lambda(t) = \kappa(t + c)^{-p}$, ha sido popularmente adoptada por los analistas para explicar la frecuencia de réplicas de terremotos por unidad de tiempo. En dicha expresión t es el lapso de tiempo a partir del terremoto principal (suponiendo que el origen $t = 0$, corresponde al tiempo de ocurrencia del terremoto principal) y κ, c y p son parámetros que normalmente se estiman por máxima verosimilitud. Estos parámetros reflejan condiciones físicas de la corteza terrestre. Se observa que la actividad de las réplicas decae rápidamente en regiones con mayor temperatura en la corteza terrestre, ver Ogata (1999) para una discusión más detallada.

Seguidamente, se analizan dos conjuntos de datos relativos a réplicas de terremotos. El primer conjunto de datos consta de 2305 tiempos (en días) de la ocurrencia de réplicas que suceden días después del terremoto principal ocurrido en Miyagi-Ken (Japón) el 26 de julio del año 2003 con una magnitud de 6.2 en la escala de Richter. Los datos están disponibles en el libro SAPP de R. Con esta información se construye el SiZer Map para explorar las características que subyacen a la función de intensidad del proceso. En la Figura 3.13 se muestra el Family Plot en la gráfica superior y los resultados de la inferencia sobre la derivada en la gráfica inferior, o sea en el SiZer Map. Se consideran 11 parámetros de suavizado permitiendo una visualización apropiada del problema y una red con 401 tiempos comprendidos dentro del rango de los datos, desde cero (día en el que sucede el terremoto principal) hasta la última ocurrencia el 28 de agosto de 2003. El Family

Plot muestra una estimación de la intensidad utilizando el estimador lineal local con cada parámetro de suavizado (líneas azules), así como la función paramétrica de Omori (línea discontinua) descrita anteriormente, ajustada mediante máxima verosimilitud utilizando la función `omori()` del libro SAPP de R. En la Figura 3.13, puede verse que la intensidad muestra un rápido decrecimiento después del origen pero tal vez no tan marcado como el que se describe mediante la ley de Omori. Por otro lado parece que para algunos parámetros de suavizado la intensidad muestra pocos cambios de tendencia, siendo el más notable el que se encuentra alrededor del quinto día después del terremoto principal. Para confirmar si estas características son en realidad significativas, en el SiZer Map se observan los cambios significativos del signo de la primera derivada utilizando el código de colores. En esta figura se han considerado los intervalos de confianza simultáneos calculados a partir de la aproximación Normal. Similares resultados se han obtenido con los intervalos de confianza calculados a partir de los demás cuantiles considerados. El SiZer Map muestra un decrecimiento significativo después del terremoto principal (color rojo) y un pico significativo alrededor del quinto día después del terremoto principal (cambio de color de azul a rojo). Esto significa que, para este conjunto de datos, el modelo de Omori parece subestimar el riesgo después del terremoto principal, ignorando un crecimiento significativo alrededor del quinto día.

El segundo conjunto de datos consta de 150 tiempos de observación de réplicas (en días) después del terremoto principal sucedido en Sichuan (China) el 12 de mayo del año 2008 con una magnitud de 7.9 en la escala de Richter. Estos datos pueden encontrarse en la United States Geological Survey (USGS) en `https://es.wikipedia.org/wiki/Terremoto_de_Sichuan_de_2008`. El conjunto de datos se restringe hasta el 7 de agosto de 2008 ya que las ocurrencias de réplicas eran muy dispersas y los análisis se vuelven poco fiables. El análisis SiZer para estos datos viene dado en la Figura 3.14. En el Family Plot la línea discontinua muestra el modelo de Omori estimado por máxima verosimilitud. El SiZer Map

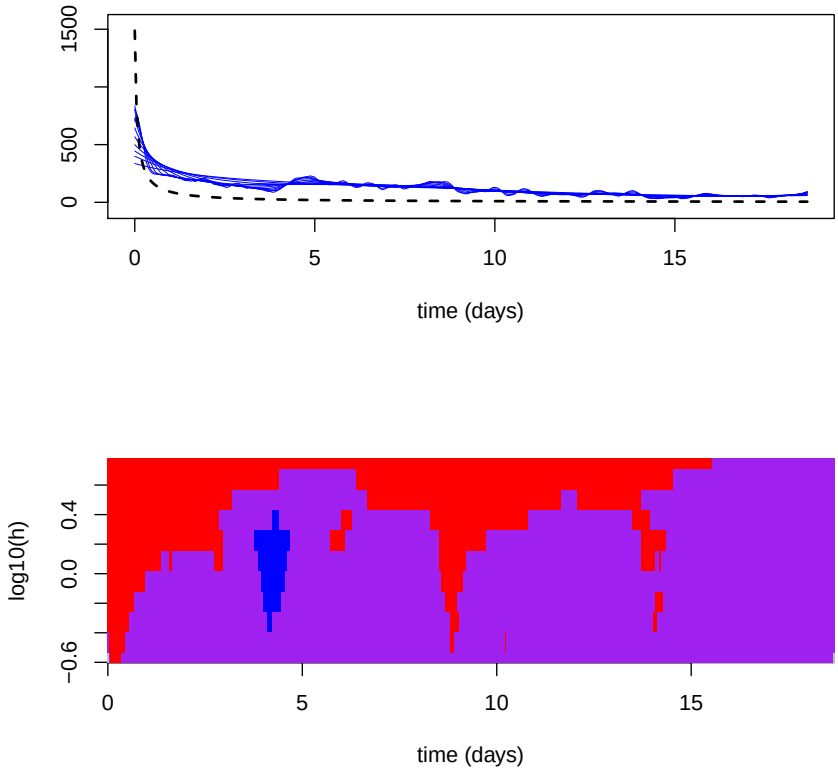


Figura 3.13: Análisis SiZer con intervalos de confianza Normal de tipo simultáneo para datos de réplicas del terremoto de magnitud 6.2 en la escala de Richter sucedido el 26 de julio de 2003 en el norte de Miyagi-Ken (Japón).

se ha construido utilizando intervalos de confianza simultáneos de tipo bootstrap. Puede apreciarse que, en este caso, la intensidad se describe bastante bien con la fórmula de Omori, ya que no se destacan características significativas por el SiZer Map después del terremoto principal y el decaimiento es muy rápido. Mediante SiZer Map se confirma que la actividad sísmica decrece rápidamente en los días posteriores al terremoto principal, ya que en dicha gráfica el color rojo significa que el intervalo de la derivada es negativo, y conforme avanza en el tiempo se vuelve a color púrpura. Esta gráfica se ha construido con intervalos de confianza simul-

táneos de tipo bootstrap a partir de 1000 muestras bootstrap y las salidas con los diferentes intervalos de confianza considerados aquí son muy parecidas. Chen *et al.* (2011) muestra conclusiones similares analizando este conjunto de datos con el estimador lineal local de la derivada. Sin embargo sus conclusiones son altamente dependientes del parámetro de suavizado elegido para su estimador y además no proporciona inferencia para contrastar la significación. El análisis gráfico proporcionado por SiZer Map aclara y confirma estas conclusiones.

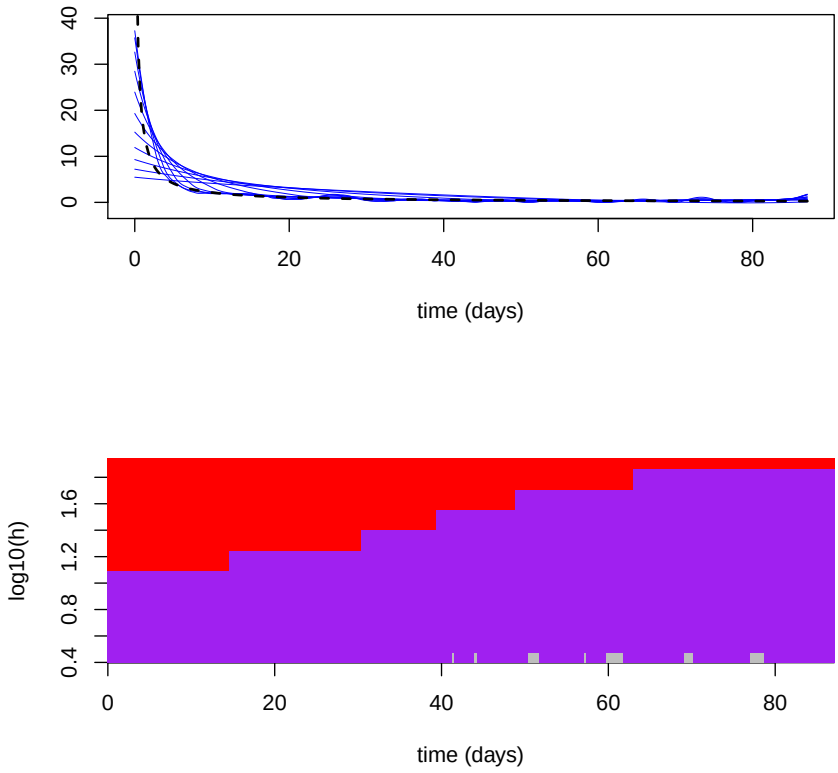


Figura 3.14: Análisis SiZer con intervalos de confianza Normales de tipo simultáneo para datos de réplicas del terremoto de magnitud 7.9 en la escala de Richter sucedido el 12 de mayo de 2008 en Sichuan (China).

3.6.2. Datos de un sistema hidráulico de máquinas de carga, acarreo y descarga

En este ejemplo, se describen unos datos de Fiabilidad. Los datos han sido tomados de Kumar y Klefsjö (1992) y consisten en tiempos entre fallos sucesivos (en horas, excluyendo reparaciones o tiempos de paradas) de sistemas hidráulicos de seis máquinas de carga, acarreo y descarga (Load, Haul, Dump machine, (LHD)). El tiempo de fallo de cada máquina puede ser adecuadamente representado por un PPNH. Dado que las máquinas operan de forma independiente a lo largo del tiempo, si todos los fallos ocurriesen en todas las máquinas, entonces este caso se podría considerar como un proceso general de fallo y este esquema sería un PPNH. El interés de este ejemplo surge de la necesidad de conocer posibles patrones de fallos en los sistemas hidráulicos de estás máquinas. El análisis SiZer para estos datos viene reflejado en la Figura 3.15, para construirlo, se han considerado 11 parámetros de suavizado permitiendo una visualización apropiada del problema y una red con 101 tiempos equiespaciados comprendidos dentro del rango de los datos. En este caso se consideran intervalos simultáneos construidos por el método bootstrap (con 1000 muestras bootstrap). El método bootstrap parece ser apropiado para estos datos ya que el tamaño muestral es pequeño con 151 observaciones. En la Figura 3.15, pueden apreciarse dos gráficas, la gráfica superior, es el Family Plot y representa a la familia de curvas suavizadas, esto es, conjunto de funciones de intensidad, construidas mediante los diferentes parámetros de suavizado considerados. Como puede verse, el conjunto de curvas ROCOF suavizadas tiene un comportamiento creciente hasta las 1800 horas aproximadamente y seguidamente vuelven a decrecer. La gráfica inferior es el SiZer Map que representa la inferencia sobre la derivada de las funciones ROCOF suavizadas, el cambio de color azul-púrpura-rojo alrededor de las 2000 horas significa que de manera general, las curvas ROCOF muestran un cambio de tendencia significativo en ese instante. Las características de las curvas ROCOF pueden verse claramente para todas las escalas (parámetros de suavizado)

consideradas en la gráfica.

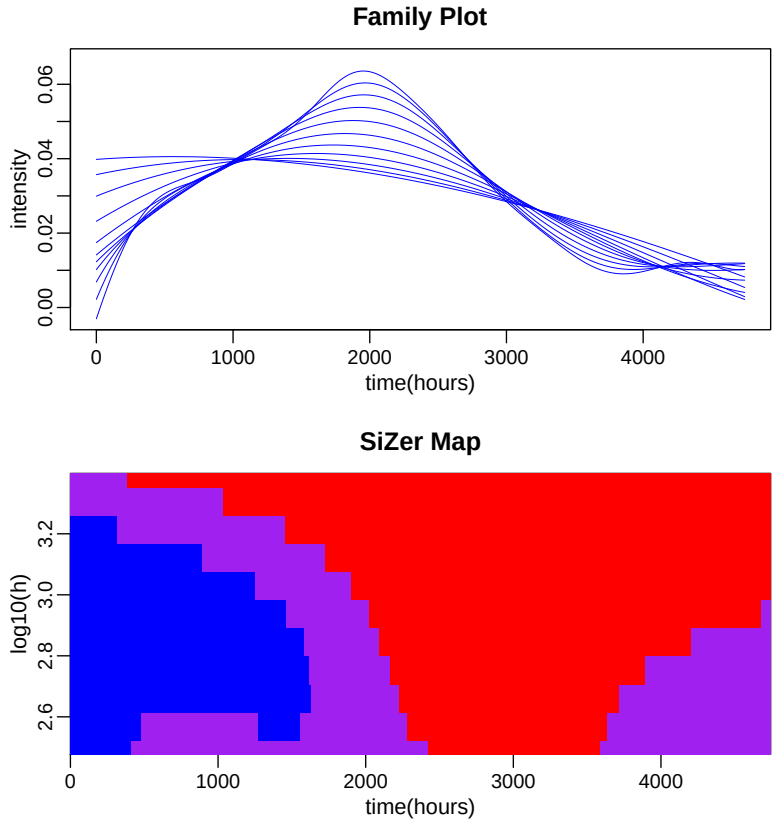


Figura 3.15: Análisis SiZer con intervalos de confianza bootstrap de tipo simultáneo para datos de un sistema hidráulico.

3.6.3. Datos del sistema del tren de potencia de un autobús urbano

En este ejemplo se describen nuevamente unos datos de Fiabilidad. Los datos han sido tomados de Guida y Pulcini (2009) y consisten en 55 tiempos de fallo (medidos en kilómetros) del sistema del tren de potencia de un autobús urbano construido por Breda Menarinibus y mantenido en servicio desde los primeros meses de 1999 hasta finales de diciembre de 2004. El autobús estuvo circulando en

las vías urbanas de la ciudad de Nápoles. El tren de potencia del autobús consta de un motor FIAT, una transmisión con caja de cambios ZF, ejes de transmisión y diferenciales. El sistema del tren de potencia estuvo sometido a reparación mínima después del fallo, debido a tres condicionantes: a) la complejidad del sistema, b) la presencia esperada de fallos prematuros, y c) el gran número de fallos/reparaciones experimentadas durante todo el periodo de observación. El interés de este ejemplo surge de la necesidad de conocer un posible patrón de fallo del sistema del tren de potencia de dichos autobuses. El análisis SiZer para estos datos viene reflejado en la Figura 3.16. De nuevo, se han considerado 11 parámetros de suavizado permitiendo una visualización apropiada del problema y una red con 41 tiempos comprendidos dentro del rango de los datos. En este caso, se consideran intervalos de confianza de tipo simultáneo construidos mediante la aproximación bootstrap. El gráfico superior de la Figura 3.16, es el Family Plot en el que puede verse que la familia de curvas ROCOF es decreciente hasta los primeros 100 kilómetros y después empieza a crecer hasta pasados los 200 kilómetros en los que vuelve a decrecer sobre los 300 kilómetros y nuevamente vuelve a crecer. En la gráfica inferior de la Figura 3.16 se aprecia que estos cambios son significativos ya que los patrones de color rojo-púrpura-azul alrededor de los 100 y los 300 kilómetros así lo indican, al igual que los dos crecimientos azul-púrpura-rojo sobre los 200 kilómetros y el azul-púrpura del final del gráfico sobre los 400 kilómetros. Por tanto, podría apreciarse un patrón de intensidades de fallos, en aumento cada 200 kilómetros. Cabe destacar que el análisis SiZer detecta un pico significativo en las curvas ROCOF alrededor de los 200 kilómetros que no detecta la aproximación paramétrica desarrollada por Guida y Pulcini (2009).

3.6.4. Datos de temporales sucedidos en el mar Ártico

En este ejemplo se consideran unos datos relativos a la ocurrencia de tormentas en el mar Ártico. Los datos han sido tomados de Lee *et al.* (1991) y consisten

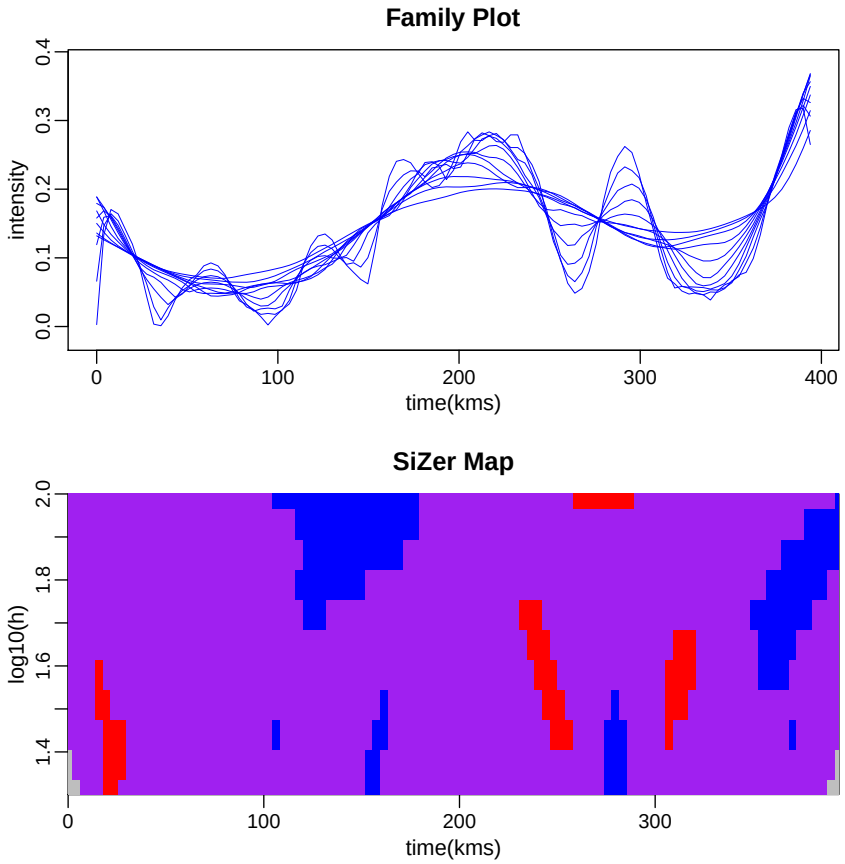


Figura 3.16: Análisis SiZer con intervalos de confianza bootstrap de tipo simultáneo para datos de un sistema de tren de potencia de un autobús urbano.

en temporales marítimos observados mensualmente durante un intervalo de nueve años consecutivos en alta mar. Durante este periodo de tiempo se observan un total de 302 temporales. De acuerdo a Lee *et al.* (1991) las series de tiempos entre tormentas suceden de manera periódica y la tasa de ocurrencia de dichos eventos varía de manera suave año tras año.

Los autores llegan a la conclusión a través de sus análisis, de que existen fuertes evidencias de que el proceso de ocurrencias de temporales es un PPNH. En consecuencia proponen un modelo paramétrico para la tasa de ocurrencia en términos de una función trigonométrica que proporciona una buena explicación del suceso

de eventos.

El análisis SiZer para estos datos viene reflejado en la Figura 3.17. Para construirlo, se han considerado 11 parámetros de suavizado permitiendo una visualización apropiada del problema y una red con 401 tiempos equiespaciados comprendidos dentro del rango de los datos. En este caso se han considerado intervalos puntuales construidos mediante la aproximación Normal. Viendo el Family Plot, se aprecia que la familia de curvas de intensidad del proceso es cíclica durante todos los años considerados, coincidiendo una mayor intensidad de ocurrencia de temporales a mitad de cada año aproximadamente. El SiZer Map muestra la significación de todos los picos que forman la familia de curvas suavizadas, mediante los cambios de color azul-púrpura-rojo alrededor de la mitad de cada año (un pico por año). El análisis SiZer, corrobora las conclusiones de Lee *et al.* (1991).

3.6.5. Datos de inmigración ilegal

En este ejemplo se consideran unos datos relativos a la inmigración ilegal sucedida en el sureste peninsular a través del mar. Los datos constan de 291 tiempos de ocurrencia (en días) relativos al rescate de pateras sucedidos en Andalucía oriental, en el periodo comprendido entre los años 2012 y 2015. Las pateras que parten del norte de África hacia la Península Ibérica, realizan la travesía intentando atravesar el Mar de Alborán. En este trayecto, los inmigrantes se encuentran con el problema de que sus pequeñas embarcaciones no están preparadas para tal viaje y quedan a la deriva, peligrando seriamente la vida de estas personas a causa de las inclemencias del mar. El rescate es realizado por la flota marítima y los medios aéreos de Salvamento Marítimo, en colaboración con la Cruz Roja y la Guardia Civil. Las embarcaciones de rescate del sureste español están ubicadas en un territorio que abarca desde Garrucha (Almería) hasta Motril (Granada), el radio de acción de estas embarcaciones cubre toda la parte centro-este del mar de Alborán y el Golfo de Vera. La motivación de este ejemplo es el estudio de la intensidad del proceso a

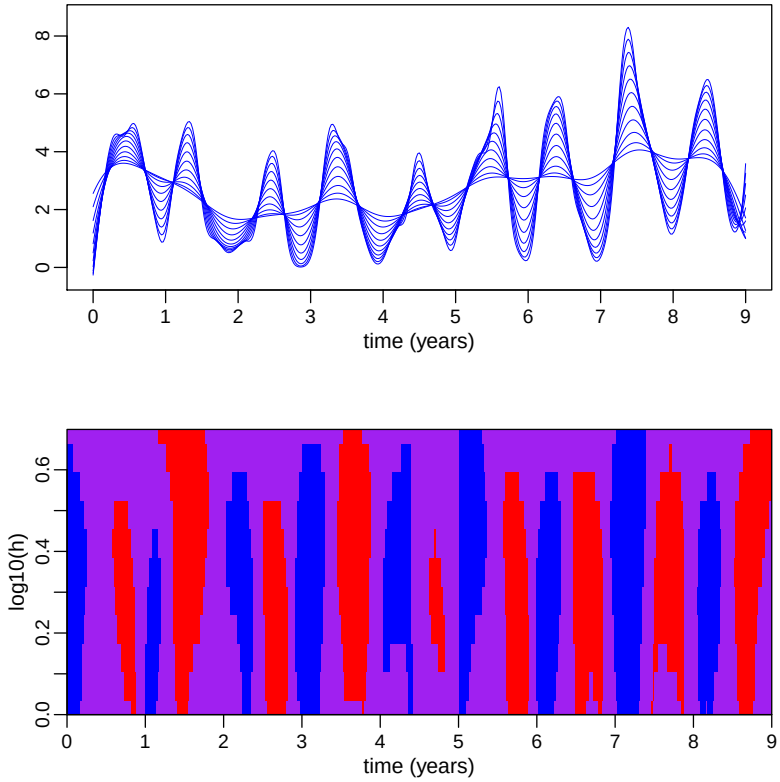


Figura 3.17: Análisis SiZer con intervalos de confianza Normal de tipo puntual para datos relativos a las tormentas sucedidas en el mar Ártico.

lo largo del tiempo, para que la empresa pueda considerar la decisión de reforzar las tres embarcaciones de rescate de intervención rápida (Salvamares) que operan en el sureste español, con un cuarto tripulante. El análisis SiZer para estos datos viene reflejado en la Figura 3.18. Para ello se han considerado 11 parámetros de suavizado permitiendo una visualización apropiada del problema y una red con 201 tiempos equiespaciados comprendidos dentro del rango de los datos. En este caso se consideran intervalos de tipo puntual construidos mediante la aproximación Normal. Viendo el Family Plot, se aprecia que la familia de curvas de intensidad es generalmente creciente durante todo el proceso, aunque pueden apreciarse unos

crecimientos y decrecimientos más acentuados, en torno a los tiempos 200, 600 y 1000 para la mayoría de las curvas suavizadas. Viendo el SiZer, se aprecian unos cambios de color azul-púrpura-rojo poco después de los 200, 600 y 1000 días, que significan que la función de intensidad muestra un cambio de tendencia significativo en ese instante, donde hay un crecimiento y decrecimiento en la curva de intensidad, que corresponde a los meses de verano de los años 2012, 2013 y 2014, respectivamente, seguido de un patrón rojo-púrpura-azul alrededor de los 400 y 800 días, que indica que después del decrecimiento anterior, la curva vuelve a crecer aproximadamente a principios del año 2013 y 2014, respectivamente. Si se analiza el estudio segmentado por años, las conclusiones no varían. En los SiZer de la Figura 3.19, puede verse que en los gráficos correspondientes a los años 2014 y 2015, se aprecia claramente una tendencia creciente en todo el gráfico, marcada por el color azul. Por todo lo demás la mayoría de las curvas de intensidad estimadas bajo diferentes parámetros de suavizado tienen un comportamiento creciente. Las características de la curva de intensidad pueden verse claramente para todas las escalas (parámetros de suavizado) consideradas en la gráfica. En resumen, se observa un comportamiento claramente creciente en la intensidad del proceso, por lo que la empresa debería considerar seriamente la idea de reforzar las embarcaciones con un cuarto tripulante en las Salvamares que no lo tengan, como sucede en la Salvamar Algenib (Garrucha) y mantener firmemente el que poseen en la Salvamar Denébola (Almería) y en la Salvamar Hamal (Motril), ya que, como se ha visto, la intensidad del proceso va en aumento en estos últimos años.

3.6.6. Datos de accidentes en minas de carbón

En este ejemplo se consideran unos datos relativos a accidentes mineros (explosiones) causados por gas grisú o por polvo de carbón que causaron la muerte de más de 10 mineros. Los datos son tomados de Jarrett (1984) y constan de 191 tiempos de ocurrencias de accidentes (en días) ocurridos durante un periodo comprendido

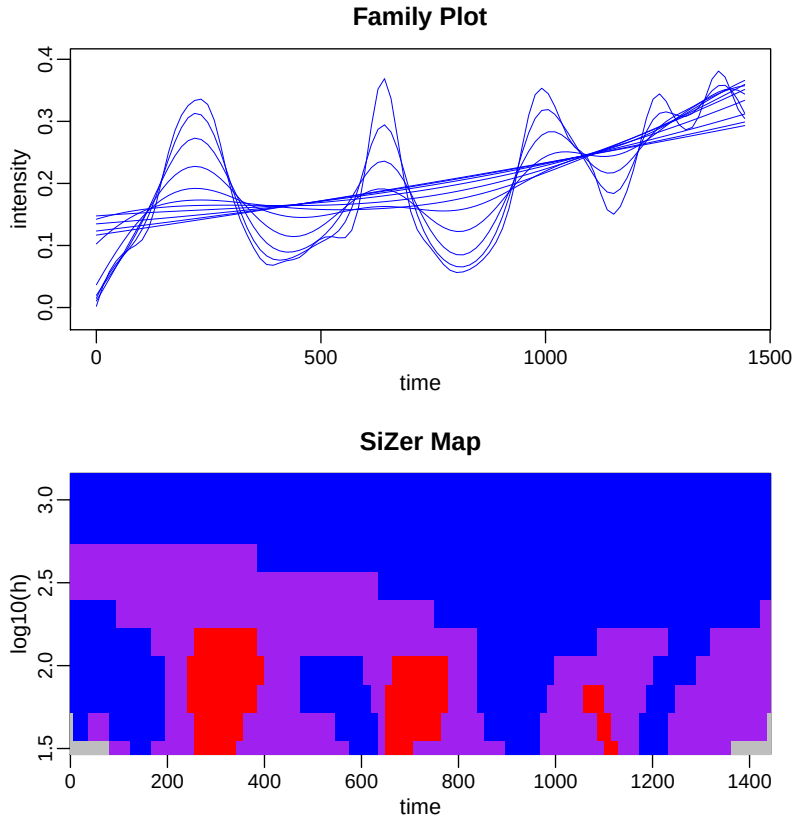


Figura 3.18: Análisis SiZer con intervalos de confianza Normal de tipo simultáneo para datos relativos al rescate de pateras en aguas españolas.

entre el 15 de marzo de 1851 y el 22 de marzo de 1962. Las ocurrencias de dichos accidentes pueden considerarse como un PPNH. El interés de este ejemplo surge de la necesidad de conocer posibles patrones de ocurrencia de dichos accidentes para una mejor comprensión en la naturaleza de estos eventos. El análisis SiZer para estos datos viene reflejado en la Figura 3.20. Se consideran 11 parámetros de suavizado que permiten una visualización apropiada del problema, una red con 101 tiempos comprendidos dentro del rango de los datos y como función núcleo, se elige como en todos los casos anteriores el núcleo de Epanechnikov. En este caso se consideran intervalos de confianza puntuales construidos mediante la aproximación

Normal. En el gráfico SiZer se aprecia que las curvas ROCOF son decrecientes hasta aproximadamente el año 1920 en donde aparece un cambio significativo ya que el patrón de color rojo-púrpura-azul, así lo indica. También se aprecia un cambio significativo, esta vez creciente, reflejado por el patrón de colores azul-púrpura-rojo alrededor del año 1940.

3.6.7. Un análisis diagnóstico de la hipótesis de PPNH en los datos reales

Esta sección trata la cuestión de detectar la posible violación de la hipótesis de que los tiempos de ocurrencia de eventos en los ejemplos considerados anteriormente, son ciertamente un PPNH, para ello se utiliza un método gráfico sugerido en Gámiz y Lindqvist (2016). El método se basa en la idea de que si el proceso de ocurrencia es un PPNH, la transformación de los tiempos de eventos sucesivos mediante la función de intensidad acumulada proporcionará tiempos de llegadas sucesivos de acuerdo con un PPNH de tasa 1 en una escala transformada. Para calcular los tiempos entre llegadas en la escala transformada, se calcula previamente un estimador no paramétrico de la función de intensidad utilizando el procedimiento presentado por Gámiz y Lindqvist (2016). Si el proceso subyacente es un PPNH, estos tiempos son i.i.d. según una ley exponencial de tasa 1 y por tanto, el riesgo estimado correspondiente debería distribuirse aproximadamente sobre la recta $y = 1$.

El algoritmo para realizar los gráficos es el siguiente

- Se considera $\widehat{\Lambda}(t) = \int_0^t \widehat{\lambda}(u)du$ un estimador de $\Lambda(t)$, siendo $\widehat{\lambda}(u)$ el estimador obtenido en la expresión (3.1).

- Se transforman los tiempos de fallo observados mediante el estimador anterior $\{\widehat{\Lambda}(T_0), \widehat{\Lambda}(T_1), \ldots, \widehat{\Lambda}(T_n)\}$.

- Se calcula la secuencia de tiempos entre llegadas de la forma $X_i = \widehat{\Lambda}(T_i) -$

$\widehat{\Lambda}(T_{i-1})$ con $i = 1, 2, \ldots, n$.

- Bajo un PPNH los tiempos $\{X_1, X_2, \ldots, X_n\}$ son i.i.d. según una ley exponencial de tasa 1.
- Se estima la función de intensidad de $\{X_1, X_2, \ldots, X_n\}$.

Para más detalles del procedimiento ver Gámiz y Lindqvist (2016).

Seguidamente, se aplica dicho procedimiento a algunos de los datos reales analizados anteriormente, asumiendo que se pueden modelizar como PPNHs. Los resultados se muestran en la Figura 3.21. En dicha figura se representa la tasa de riesgo basada en los tiempos entre llegadas en la escala transformada, para los ejemplos considerados. A la vista de estas estimaciones se puede concluir que los ejemplos escogidos se pueden modelizar de forma adecuada mediante un PPNH, ya que en dichas gráficas, los tiempos oscilan cercanos a la recta $y = 1$.

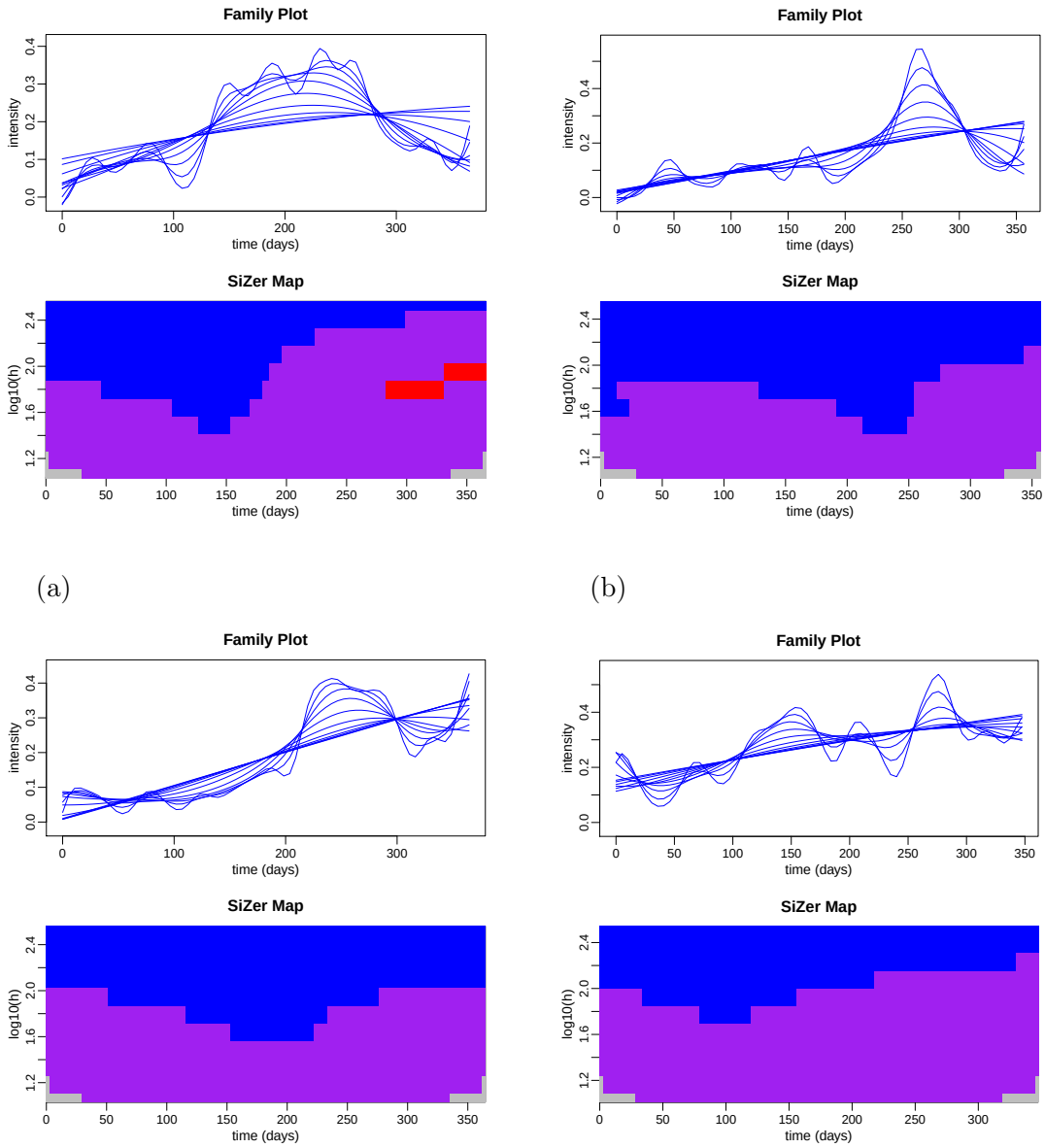


Figura 3.19: Análisis SiZer con intervalos de confianza bootstrap de tipo simultáneo para datos relativos al rescate de pateras en aguas españolas, realizados separadamente por años, siendo los años (a)= 2012, (b)= 2013, (c)= 2014 y (d)= 2015.

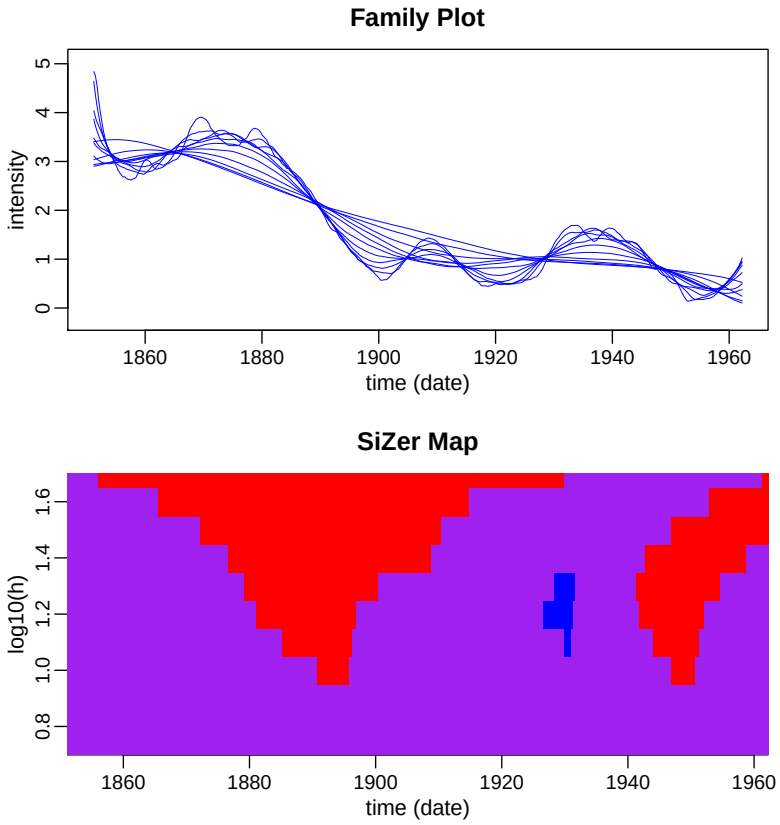


Figura 3.20: Análisis SiZer con intervalos de confianza Normal de tipo puntual para datos relativos a los accidentes ocurridos en minas de carbón.

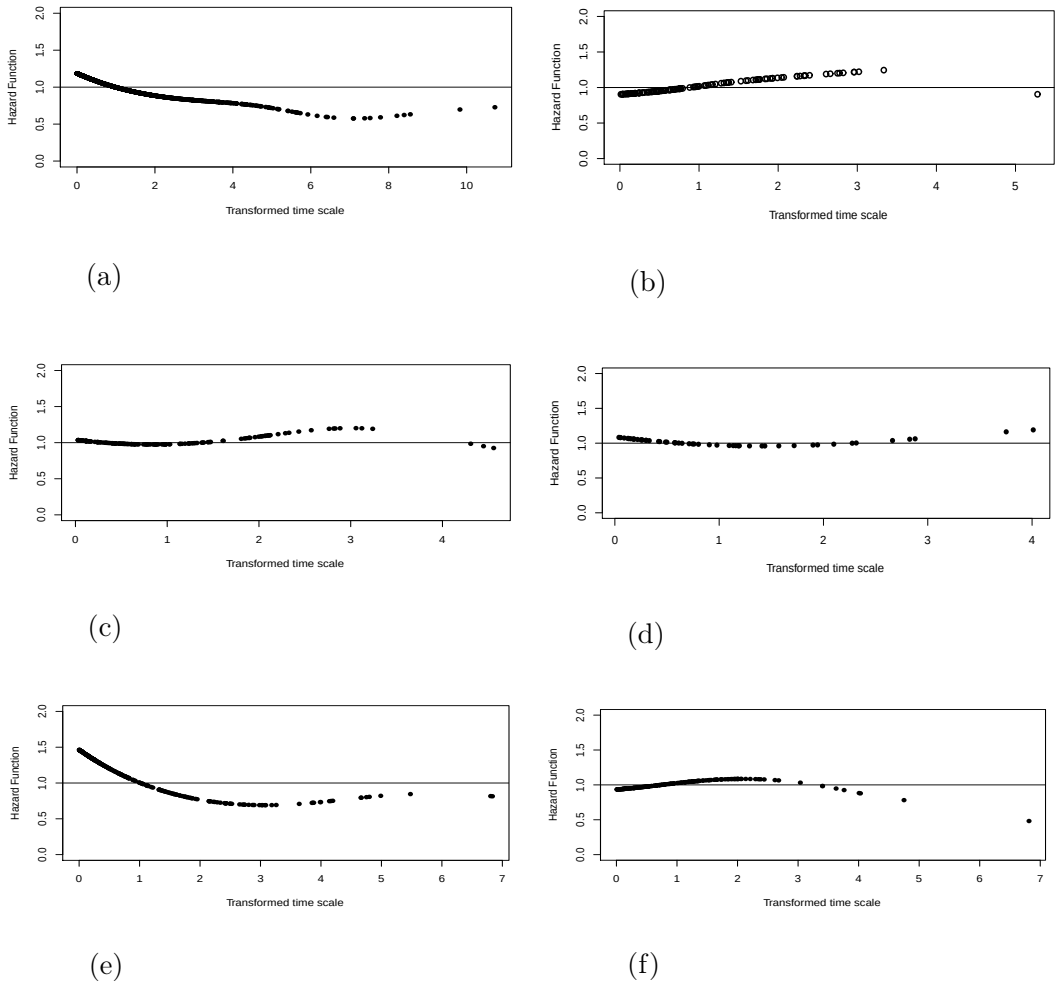


Figura 3.21: Diagnóstico gráfico de la hipótesis de PPNH para los conjuntos de datos analizados siendo (a) terremotos en Japón, (b) terremotos en China, (c) sistema hidráulico, (d) tren de potencia de un autobus urbano, (e) rescate de pateras y (f) accidentes en minas de carbón.

3.7. Resultados y conclusiones

El objetivo principal de este capítulo es la inferencia no paramétrica de la función de intensidad a través de la herramienta gráfica SiZer. A diferencia de la inferencia no paramétrica tradicional que se basa en la elección de un parámetro de suavizado óptimo para derivar las conclusiones inferenciales, el análisis mediante SiZer considera simultáneamente un amplio rango de valores del parámetro de suavizado. Esto es una aproximación muy útil desde el punto de vista exploratorio ya que diferentes niveles de suavizado pueden revelar diferentes características y estructuras en los datos.

Se analizan varios conjuntos de datos mediante SiZer y dicha herramienta demuestra ser muy útil a la hora de detectar estructuras subyacentes en los cambios de tendencia de la función de intensidad. El desarrollo del análisis de datos reales relativos a las réplicas de terremotos muestra que los modelos paramétricos (por ejemplo la fórmula de Omori) pueden ser revisados utilizando el análisis exploratorio SiZer a la vez que pueden encontrarse nuevas estructuras en los datos. Para los conjuntos de datos del ámbito de la Fiabilidad, los análisis SiZer muestran características importantes relativas a los fallos ocurridos en los sistemas considerados, que serán relevantes para las posibles mejoras que los fabricantes realicen sobre dichas máquinas. Para los conjuntos de datos relativos a los temporales sucedidos en el Ártico, el análisis SiZer sirve de soporte a estudios previos realizados sobre la intensidad de dichos fenómenos. Para los conjuntos de datos sobre las emergencias asistidas, se ofrece información a la empresa E.P.E. Sociedad de Salvamento y Seguridad Marítima en su toma de decisiones sobre inmigración ilegal.

El extenso estudio de simulación realizado en este capítulo confirma la eficacia de SiZer Map a la hora de detectar cambios de tendencia significativos en la intensidad.

Una extensión natural de la metodología propuesta en este capítulo, consiste en considerar otros escenarios más generales donde la función de intensidad no es

una función determinística. Este caso constituye una futura línea de investigación.

Bibliografía

[1] Aalen O.O. (1975). *Statistical inference for a family of counting processes*. PhD thesis, University of California, Berkley.

[2] Aalen O.O. (1978). *Nonparametric Inference for a Family of Counting Processes*. The Annals of Statistics, **6**, 701 – 726.

[3] Andersen P.K., Borgan O., Gill R.D. and Keiding N. (1993). *Statistical Models Based on Counting Processes*. Springer Series in Statistics, New York.

[4] Borgan O. (1984). *Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data*. Scandinavian Journal of Statistics, **11**, 1 – 16.

[5] Bowman A.W. (1984). *An alternative method of cross-validation for the smoothing of density estimates*. Biometrika, **71**, 353 – 360.

[6] Breslow N.E. (1972). *"Discussion on Professor Cox's paper", Introduction to Stochastic Processes in Biostatistics*. Communications in Statistics - Simulation and Computation, John Wiley and Sons, New York.

[7] Brooks M.M. and Marron J.S. (1991). *Asymptotic optimality of the least-squares cross-validation bandwidth for kernel estimates of intensity functions*. Stochastic Processes and Their Applications, **38**, 157 – 165.

[8] Broström G. (2012). *Event History Analysis with R*. Chapman & Hall/CRC The R Series, United States.

[9] Buckley J.J. and James I.R. (1979). *Linear regression with censored data.* Biometrika, **66**, 429 – 436.

[10] Carrión A., Solano H., Gámiz M.L. and Debón A. (2010). *Evaluation of the reliability of a water supply network from right-censored and left-truncated break data.* Water Resources Management, **24**, 2917 – 2935.

[11] Chaudhuri P. and Marron J.S. (1999). *SiZer for exploration of structure in curves.* Journal of the American Statistical Association, **94**, 807 – 823.

[12] Chen F., Huggins R.M., Yip P.S.F. and Lam K.F. (2008). *Nonparametric estimation of multiplicative counting process intensity functions with an application to the Beijing SARS epidemic.* Communications in Statistics - Theory and Methods, **37**, 294 – 306.

[13] Chen F., Yip P.S.F. and Lam K.F. (2011). *On the Local Polynomial Estimators of the Counting Process Intensity Function and its Derivatives.* Scandinavian Journal of Statistics, **38**, 631 – 649.

[14] Courrège P. and Priouret P. (1965). *Temps d'arrêt d'un fonction aléatoire.* Publications de l'Institut de statistique de l'Université de Paris.

[15] Cowling A., Hall P. and Phillips M.J. (1996). *Bootstrap confidence regions for the intensity of a Poisson point process.* Journal of the American Statistical Association, **91(436)**, 1516 – 1524.

[16] Cox D.R. (1972). *Regression models and life-tables (with discusion).* Journal of the Royal Statistical Society: Series B (Statistical Methodology), **34**, 187 – 220.

[17] Diggle P. and Marron J.S. (1988). *Equivalence of smoothing parameter selectors in density and intensity estimation.* JASA, **83**, 793 – 800.

[18] Efron B. and Tibshirani R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

[19] Fan J. and Gijbels I. (1996). *Local polynomial modelling and its applications*. Chapman & Hall, United States.

[20] Fernholz L.T. (1991). *Almost sure convergence of smoothed empirical distribution functions*. Scandinavian Journal of Statistics, **18**, 225 – 262.

[21] Fleming T. and Harrington D. (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons, New York.

[22] Gámiz M.L., Kulasekera K.B., Limnios N. and Lindquist B.H. (2011). *Applied Nonparametric Statistic in Reliability*. Springer Series in Reliability Engineering, New York.

[23] Gámiz-Pérez M.L., Martínez-Miranda M.D. and Nielsen J.P. (2013). *Smoothing survival densities in practice*. Computational Statistics and Data Analysis, **58**, 368 – 382.

[24] Gámiz-Pérez M.L., Janys L., Martínez-Miranda M.D. and Nielsen J.P. (2013). *Bandwidth selection in marker dependent kernel hazard estimation*. Computational Statistics and Data Analysis, **68**, 155 – 169.

[25] Gámiz M.L. and Lindqvist B.H. (2016). *Nonparametric estimation in trend-renewal processes*. Reliability Engineering and System Safety, **145**, 38 – 46.

[26] Gámiz M.L., López-Montoya A.J., Martínez-Miranda M.D. y Raya-Miranda R. (2017). *Visual inference for NHPP intensities*. (Sometido).

[27] Gasser T. and Müller H.G. (1979). *Kernel estimation of regression functions. In: Lecture Notes in Mathematics, 757, pp. 23-68*. Springer, New York.

[28] Godtliebsen F., Marron J.S. and Chaudhuri P. (2002). *Significance in scale space for bivariate density estimation.* Journal of Computational and Graphical Statistics, **11**, 1 – 21.

[29] González-Manteiga W., Martínez-Miranda M.D. and Raya-Miranda R. (2008). *SiZer Map for inference with additive models.* Statistics and Computing, **18**, 297 – 312.

[30] Gross S.T. and Lai T.L. (1996). *Bootstrap Methods for Truncated and Censored Data.* Statistica Sinica, **6**, 509 – 530.

[31] Guida M. and Pulcini G. (2009). *Reliability analysis of mechanical systems with bounded and bathtub-shaped intensity function.* IEEE Transactions on Reliability, **58(3)**, 432 – 443.

[32] Hannig J. and Lee T.C.M. (2006). *Robust SiZer for exploration of regression structures and outlier detection.* Journal of Computational and Graphical Statistics, **15**, 101 – 117.

[33] Jarrett R.G. (1984). *A note on the intervals between coal-mining disasters.* Biometrika, **66**, 191 – 193.

[34] Jin Z., Lin D., Wei L.J. and Ying Z. (2003). *Rank-based inference for the accelerated failure time model.* Biometrika, **90**, 341 – 353.

[35] Jones M.C. (1993). *Simple boundary correction for kernel density estimation.* Statistics and Computing, **3**, 135 – 146.

[36] Jones M.C., Davies S.J. and Park B.U. (1994). *Versions of kernels-type regression estimator.* Journal of the American Statistical Association, **89**, 825 – 832.

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

[38] Kaplan E.L. and Meier P. (1958). *Nonparametric estimation from incomplete observations*. Journal of the American Statistical Association, **53**, 457 – 481.

[39] Klein J.P. and Moeschberger M.L. (2003). *Survival analysis: Techniques for censored and truncated data*. Springer, New York.

[40] Kleinbaum D.G. and Klein M. (2005). *Survival analysis: a self-learning text*. Springer, New York.

[41] Krivtsov V.V. (2007). *Practical extensions to NHPP application in repairable system reliability analysis*. Reliability Engineering and System Safety, **92(5)**, 560 – 562.

[42] Kulasekera K.B., Williams C.L., Coffin M. and Manatunga A. (2001). *Smooth Estimation of the Reliability Function*. Lifetime Data Analysis, **7**, 413 – 433.

[43] Kumar U. and Klefsjö B. (1992). *Reliability analysis of hydraulic system of LHD machines using the power-law process model*. Reliability Engineering and System Safety, **35**, 217 – 224.

[44] Lawless J.F. (1982). *Statistical Models and Methods for Lifetime Data Analysis*. John Wiley & Sons, New York.

[45] Lai T.L. and Ying Z. (1991). *Rank Regression Methods for Left-Truncated and Right-Censored Data*. Annals of Statistics, **19**, 505 – 1108.

[46] Leadbetter M.R. and Wold D. (1983). *On estimation of point process intensities. In: Contributions to Statistics: Essays in Honor of Norman L. Johnson*. 299-312, ed. P.K. Sen, Amsterdam: North Holland.

[47] Lee S., Wilson J.R. and Crawford M.M. (1991). *Modeling and simulation of a nonhomogeneous Poisson process having cyclic behavior*. Communications in Statistics - Simulation and Computation, **20(2&3)**, 777 – 809.

[48] Li R. and Marron J.S. (2005). *Local likelihood SiZer map*. Sankhya: The Indian Journal of Statistics, **67(3)**, 476 – 498.

[49] Lindeberg T. (1994). *Scale-Space Theory in Computer Vision*. The Springer International Series in Engineering and Computer Science, Kluwer, Boston.

[50] López-Montoya A.J., Gámiz-Pérez M.L. and Martínez-Miranda M.D. (2015). *Local linear smoothing to estimate accelerated lifetime model with censoring and truncation*. Applied Mathematical Modelling, **39**, 4630 – 4645.

[51] Marron J.S and de Uña-Ávarez J. (2004). *SiZer for length biased, censored density and hazard estimation*. Journal of Statistical Planning and Inference, **121**, 149 – 161.

[52] Martínez-Miranda M.D., Raya-Miranda R, González-Manteiga W. and González-Carmona A. (2008). *A local bandwidth selector for additive models*. Journal of Computational and Graphical Statistics, **17**, 38 – 55.

[53] Martinussen T. and Scheike T.H. (2006). *Dynamic regression models for survival data*. Springer, Berlin.

[54] Miller R.G. (1976). *Least squares regression with censored data*. Biometrika, **63**, 449 – 464.

[55] Müller H.G. and Wang J.L. (1994). *Hazard rate estimation under random censoring with varying kernels and bandwidths*. Biometrics, **50**, 61 – 76.

[56] Nadaraya E.A. (1964). *Some new estimates for distribution functions*. Theory of Probability and Its Applications, **15**, 497 – 500.

[57] Nelson W.B. (1969). *Hazard plotting for incomplete failure data*. Journal of Quality Technology, **1**, 27 – 52.

[58] Nelson W.B. (1972). *Theory and applications of hazards plotting for censored failure data*. Technometrics, **14**, 945 – 965.

[59] Nelson W.B. (1990). *Statistical Models and Methods for Lifetime Data Analysis*. John Wiley & Sons, New York.

[60] Nielsen J.P. and Tanggaard C. (2001). *Boundary and bias correction in kernel hazard estimation*. Scandinavian Journal of Statistics, **28**, 675 – 698.

[61] Nielsen J.P., Tanggaard C. and Jones M.C. (2009). *Local linear density estimation for filtered survival data*. Statistics, **43**, 176 – 186.

[62] Nielsen J.P. (1998a). *Marker dependent kernel hazard estimation from local linear estimation*. Scandinavian Acturial Journal, **2**, 113 – 124.

[63] Nielsen J.P. (1998b). *Multiplicative bias correction in kernel hazard estimation*. Scandinavian Journal of Statistics, **25**, 541 – 553.

[64] Nielsen J.P. and Linton O.B. (1995). *Kernel estimation in a nonparametric marker dependent hazard model*. Annals of Statistics, **23**, 1735 – 1748.

[65] Ogata Y. (1999). *Seismicity Analysis through Point-process Modeling: A Review*. Pure and Applied Geophysics, **155**, 471 – 507.

[66] Oliveira M. (2013). *Nonparametric circular methods for density and regression*. PhD thesis, Universidade de Santiago de Compostela, Spain.

[67] Oliveira M., Crujeiras R.M. and Rodríguez-Casal A. (2013). *Nonparametric circular methods for exploring environmental data*. Environmental and Ecological Statistics, **20**, 1 – 17.

[68] Orbe J., Ferreira E. and Núñez-Antón V. (2002). *Comparing proportional hazards and accelerated failure time models for survival analysis*. Statistics in Medicine, **21**, 3493 – 3510.

[69] Park C., Hannig J. and Kang K.-H. (2009). *Improved SiZer for Time Series.* Statistica Sinica, **19**, 1511 – 1530.

[70] Park C., Lee T.C.M. and Hannig J. (2010). *Multiscale exploratory analysis of regression quantiles using quantile SiZer.* Journal of Computational and Graphical Statistics, **19**, 497 – 513.

[71] Park C., Hernandez-Campos F., Le L., Marron J.S., Park J., Pipiras V., Smith F.D., Smith R.L., Trovero M. and Zhu Z. (2011). *Long Range Dependence Analysis of Internet Traffic.* Journal of Applied Statistics, **38**, 1407 – 1433.

[72] Park C., Hanning J. and Kang K.-H. (2014). *Nonparametric Comparison of Multiple Regression Curves in Scale-Space.* Journal of Computational and Graphical Statistics, **23**, 657 – 677.

[73] Parzen E. (1962). *On estimation of a probability density function and mode.* Annals of Mathematical Statistics, **33(3)**, 1065 – 1076.

[74] Phillips M.J. (2000). *Bootstrap confidence regions for the expected ROCOF of a repairable system.* IEEE Transactions on Reliability, **49(2)**, 204 – 208.

[75] Phillips M.J. (2001). *Estimation of the expected ROCOF of a repairable system with bootstrap confidence region.* Quality and Reliability Engineering International, **17**, 159 – 162.

[76] R Core Team (2015). *R Foundation for Statistical Computing. R Development Core Team. R: A Language and Environment for Statistical Computing.* Vienna, Austria, URL http://www.R-project.org.

[77] Ramlau-Hansen H. (1983). *Smoothing counting process intensities by means of kernel functions.* Annals of Statistics, **11**, 453 – 466.

[78] Rebolledo R. (1980). *Central limit theorems for local martingales.* Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, **51**, 269 – 286.

[79] Rigdon S.E. and Basu A.P. (2000). *Statistical methods for the reliability of repairable systems*. John Wiley & Sons, New York.

[80] Ritov Y. (1990). *Estimation in a linear regression model with censored data*. Annals of Statistics, **18**, 303 – 328.

[81] Rosenblatt M. (1956). *Remarks on some nonparametric estimates of a density function*. Annals of Mathematical Statistics, **33(3)**, 832 – 837.

[82] Rudemo M. (1982). *Empirical choice of histograms and kernel density estimators*. Scandinavian Journal of Statistics, **9**, 65 – 78.

[83] Ruppert D. and Wand M.P. (1994). *Multivariate locally weighted least squares regression*. Annals of Statistics, **21**, 1346 – 1370.

[84] Schoenfeld D. (1982). *Residuals for the proportional hazards regresssion model*. Biometrika, **69(1)**, 239 – 241.

[85] Silverman B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

[86] Spierdijk L. (2008). *Nonparametric conditional hazard rate estimation: a local lineal approach*. Computational Statistics & Data Analysis, **52**, 2419 – 2434.

[87] Stute W. (1993). *Consistent estimation under random censorship when covariables are present*. Journal of Multivariate Analysis, **45**, 89 – 103.

[88] Stute W. (1996a). *Distributional convergence under random censorship when covariables are present*. Scandinavian Journal of Statistics, **23**, 461 – 471.

[89] Stute W. (1996b). *The jack-knife estimate of variance of a Kaplan-Meier integral*. Annals of Statistics, **24**, 2679 – 2704.

[90] Therneau T.M. and Grambsch P.M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York.

[91] Therneau T.M. and Lumley T. (2008). *Survival analysis, including penalised likelihood*. R package version 2.34-1.

[92] Tsiatis A.A. (1990). *Estimating regression parameters using linear rank tests for censored data*. Annals of Statistics, **18**, 354 – 372.

[93] Wand M.P. and Jones M.C. (1995). *Kernel smoothing*. Chapman & Hall, London.

[94] Watson G.S. (1964). *Smooth regression analysis*. Sankhya Series A, **26**, 359 – 372.

Apéndice A

Código en R del estudio de simulación del Capítulo 2

```
##############################################################################
##############################################################################
#################### IMPLEMENTACIÓN DEL CAPÍTULO 2 ###########################
##############################################################################
##############################################################################

##############################################################################
############################ Libros necesarios ###############################
##############################################################################

library(foreign)
library(Matrix)
library(Hmisc)
library(survival)
library(eha)
library(boot)
library(bootstrap)

##############################################################################
################## Se carga la base de datos depurada ########################
##############################################################################
datos<-csv.get('FinalData.csv')


##############################################################################
##### Estimación del modelo PH de Cox con censura y truncamiento ####
##############################################################################

rc<-coxph(Surv(Truncated,Failuretime,Censor) ~ DuctileCI + GrayCI +
```

```
    Polyethylene + Sidewalk + Normal + Length + Diameter, data=datos,
    method='breslow')
summary(rc)

################################################################################
################## Hipótesis de PH del modelo de Cox ###########################
################################################################################

################################################################################
############ Gráfica del riesgo estimado para el material ######################
################################################################################
win.graph()
plot(survfit(Surv(Truncated,Failuretime,Censor) ~ Material, data=datos
    ), fun="cumhaz",
      xlab='Duración (años)', ylim=c(0, 0.65), ylab='Riesgo acumulado',
     lty=1:4,
      col=2:5, main="Función de riesgo acumulado para la covariable
    Material")
legend("topleft", legend=c("Fundición Dúctil", "Fundición
Gris", "Polietileno", "Fibrocemento"), text.width=21, cex=1.0, lty
    =1:4,
       col=c("red", "green", "blue", "cyan"))

################################################################################
############ Gráfica del riesgo estimado para el tráfico #######################
################################################################################
win.graph()
plot(survfit(Surv(Truncated,Failuretime,Censor) ~ Traffic, data=datos)
    , fun="cumhaz",
      xlab='Duración (años)', ylim=c(0, 0.80), ylab='Riesgo acumulado',
     lty=1:4,
      col=2:5, main="Función de riesgo acumulado para la covariable
    Tráfico")
legend("topleft", legend=c("Acera", "Tráfico Normal", "Tráfico Pesado"
    ), text.width=17, cex=1.0, lty=1:4,
       col=c("red", "green", "blue"))

################################################################################
############ Gráfica del riesgo estimado para la Longitud ######################
################################################################################
win.graph()
plot(survfit(Surv(Truncated,Failuretime,Censor) ~ LengthCat, data=
    datos), fun="cumhaz",
      xlab='Duración (años)', ylim=c(0, 1.0), ylab='Riesgo acumulado',
    lty=1:4,
      col=2:5, main="Función de riesgo acumulado para la covariable
    Longitud")
```

```
64 legend("topleft", legend=c("<2 m", "[2, 10) m", "[10, 50) m", ">=50 m"
       ), text.width=13, cex=1.0, lty=1:8,
65         col=c("red", "green", "blue", "cyan"))
66
67 ##############################################################################
68 ############ Gráfica del riesgo estimado para el Diámetro ##############
69 ##############################################################################
70 win.graph()
71 plot(survfit(Surv(Truncated,Failuretime,Censor) ~ DiameterCat, data=
       datos), fun="cumhaz",
72       xlab='Duración (años)', ylim=c(0, 0.6), ylab='Riesgo acumulado',
       lty=1:4,
73       col=2:5, main="Función de riesgo acumulado para la covariable
       Diámetro")
74 legend("topleft", legend=c("<=90 mm", "(90, 175] mm", "(175, 300] mm",
        ">300 mm"), text.width=13, cex=1.0, lty=1:8,
75         col=c("red", "green", "blue", "cyan"))
76
77 ##############################################################################
78 ######################### Test de hipótesis de PH ############################
79 ##############################################################################
80 rp<-cox.zph(rc)
81 print(rp)
82
83 ##############################################################################
84 ########## Estudio de los residuos del modelo de PH de Cox ############
85 ##############################################################################
86
87 ##############################################################################
88 ################## Gráfica de residuos de Cox-Snell ####################
89 ##############################################################################
90 coxsnellres<-datos$Censor-resid(rc, type="martingale")
91 rc2<-survfit(Surv(coxsnellres,datos$Censor)~1)
92 Htilde<-cumsum(rc2$n.event/rc2$n.risk)
93 win.graph()
94 plot(rc2$time, Htilde, type='s', col='blue', main=" Residuos de Cox-
       Snell",
95       xlab='Residuos de Cox-Snell modificados', ylab='Riesgo acumulado'
       )
96 abline(0,6,col='red',lty = 2)
97
98 ##############################################################################
99 ########## Gráfica de los Residuos escalados de Schoenfeld ###########
100 ##############################################################################
101 win.graph()
102 par(mfrow=c(3, 3))
103 plot(rc)
```

```
104
105 ##############################################################################
106 ############### Gráfica de los Residuos de martingala ########################
107 ##############################################################################
108 win.graph()
109 par(mfrow=c(2,2))
110 res<-residuals(rc, type=c("martingale"))
111 X<-as.matrix(datos[ ,c("Length", "Diameter")]) # matriz de covariables
112 for (j in 1:2) { # gráficas de residuos
113     plot(X[ ,j], res, xlab = c("Longitud", "Diámetro")[j], ylab="
    Residuos de Martingala")
114     abline(h=0, lty=2)
115     lines(lowess(X[ ,j], res, iter=0))
116 }
117 win.graph()
118 par(mfrow=c(2,2))
119 b <- coef(rc)[c(6,7)] # coeficientes de regresión
120 for (j in 1:2) { # gráficas de los residuos parciales
121     plot(X[ ,j], b[j]*X[ ,j] + res, xlab=c("Longitud", "Diámetro")[j],
122          ylab="Componente+Residuo")
123     abline(lm(b[j]*X[ ,j] + res ~ X[ ,j]), lty=2)
124     lines(lowess(X[ ,j], b[j]*X[ ,j] + res, iter=0))
125 }
126
127 ##############################################################################
128 ############################ Residuos  dfbeta ################################
129 ##############################################################################
130 dfbeta<-residuals(rc, type=c("dfbeta"))
131 win.graph()
132 par(mfrow=c(3,3))
133 for (j in 1:7) {
134     plot(dfbeta[ ,j], ylab=names(coef(rc))[j])
135     abline(h=0, lty=2)
136 }
137
138 ##############################################################################
139 ############## Estimación del modelo AFT Paramétrico #########################
140 ##################### con censura y truncamiento #############################
141 ##############################################################################
142
143 ##############################################################################
144 ########## Modelo de AFT con distribución de tipo Weibull ####################
145 ##############################################################################
146 aftmweib <- aftreg(Surv(Truncated,Failuretime,Censor) ~ DuctileCI +
    GrayCI + Polyethylene + Sidewalk + Normal + Length + Diameter ,
    data=datos, dist="weibull")
147 summary(aftmweib)
```

```
148
149 ##############################################################################
150 ############ Modelo AFT con distribución de tipo Lognormal ###########
151 ##############################################################################
152 aftmlogn <- aftreg(Surv(Truncated,Failuretime,Censor) ~ DuctileCI +
    GrayCI + Polyethylene + Sidewalk + Normal + Length + Diameter ,
    data=datos, dist="lognormal")
153 summary(aftmlogn)
154
155 ##############################################################################
156 ############ Modelo AFT con distribución de tipo Loglogística #########
157 ##############################################################################
158 aftmloglog <- aftreg(Surv(Truncated,Failuretime,Censor) ~ DuctileCI +
    GrayCI + Polyethylene + Sidewalk + Normal + Length + Diameter ,
    data=datos, dist="loglogistic")
159 summary(aftmloglog)
160
161 ##############################################################################
162 ############## Criterio de selección de modelos de Akaike #############
163 ##############################################################################
164 extractAIC(aftmweib)
165 extractAIC(aftmlogn)
166 extractAIC(aftmloglog)
167
168 ##############################################################################
169 ############## Estimación del modelo semi-paramétrico AFT ##############
170 ##################### con censura y truncamiento ########################
171 ##############################################################################
172
173 ##############################################################################
174 ######## Se pasan los tiempos de fallo a escala logarítmica #########
175 ##############################################################################
176 lnY<-log(datos$Failuretime)
177 lnY<-data.matrix(lnY)###se pasan a matriz
178
179 ##############################################################################
180 ############ Se implementa la función de riesgo acumulado #############
181 ##################### con censura y truncamiento ########################
182 ##############################################################################
183 datos<-data.matrix(datos)###se pasan a matriz
184 x4<-datos[ ,4] ##tiempo de fallo
185 x11<-datos[ ,11] ##código de censura
186 x13<-datos[ ,13] ##tiempo antes del 2000
187 hazcum<-function(tf,cc,tc){
188     risk<-function(x,tc,tf){sapply(1:length(x),function(i) length(
    which(tc<x[i] & x[i]<=tf)))}
189     r<-risk(x=tf, tc=tc, tf=tf)  # tamaño de riesgo fijado para cada
```

```
    elemento
190     Hi<-cumsum(cc/r) # estimador de Nelson-Aalen de la función de
    riesgo acumulado
191     return(Hi)
192 }
193 nae <- hazcum(x4, x11, x13)
194 ra <- data.matrix(nae)###se pasa a formato en forma de matriz
195
196 ##############################################################################
197 ######## Se calcula la función de distribución para este caso ########
198 ##############################################################################
199 Fdist <- (1-exp(-ra))###función de distribución
200
201 ##############################################################################
202 ########################## Se calculan los pesos ############################
203 ##############################################################################
204 n<-nrow(Fdist)
205 W<-rep(0, n)
206 W[1]<-Fdist[1]
207 PesosTCW.aft<-function(Fdist){
208     for (i in 2:n){
209         W[i] <- Fdist[i]-Fdist[i-1]
210     }
211     list(W=W)
212 }
213 TCW<-PesosTCW.aft(Fdist)
214 CW<-TCW$W #Vector de los pesos
215 ####se unen los pesos a la base de datos
216 datos<-cbind(datos, CW)
217
218 ##############################################################################
219 ###### Cálculo del error estándar mediante el Método Jackknife ######
220 ##############################################################################
221
222 ##############################################################################
223 ######## Se pasa a dataframe para trabajar con la función lm() ########
224 ##############################################################################
225 datos<-as.data.frame(datos)
226
227 ##############################################################################
228 ############## Estimador de mínimos cuadrados ponderados ###############
229 ##############################################################################
230 est.coefs<-function(data) {
231     return(lm(log(Failuretime) ~ -1 + DuctileCI + GrayCI +
    Polyethylene + Sidewalk + Normal + Length + Diameter, data=data,
    weight=CW)$coefficients)
232 }
```

```
233
234 ##############################################################################
235 #### Función auxiliar que servirá para ir eliminando una entrada ####
236 #### de la base de datos ###################################################
237 ##############################################################################
238 omitir.caso<-function(data, i) {
239     d<-dim(data) ##dimensión de la base de datos, filas y columnas
240     ## d será nulo para vectores y listas
241     if (is.null(d) || (length(d)==1)) {
242         return(data[-i])
243         ## cualquier otra base de datos tendrá por lo menos dos
    dimensiones
244     } else {
245         return(data[-i, ])
246     }
247 }
248
249 ##############################################################################
250 ################## Función que realiza el jackknife ##########################
251 ##############################################################################
252 jackknife<-function(estimator, data) {
253     ## length() trabaja con vectores y listas que no tienen atributos
    de dimensiones
254     if (is.null(dim(data))) {n<-length(data)}
255     ## para cualquier otro caso se utilizará nrow()
256     else {n<-nrow(data)}
257     ## No se sabe el número de estimator() que devolverá, por lo que
    se utiliza cbind() para poner las salidas juntas
258     ## Se empieza con una estructura vacía
259     jackknife.ests<-c()
260     ## para cada caso
261     for (omit in 1:n) {
262         ## Se repite la estimación omitiendo el caso
263         reestimate<-estimator(omitir.caso(data, omit))
264         ## Se añade la re-estimación como una nueva columna
265         jackknife.ests<-cbind(jackknife.ests, reestimate)
266     }
267     ## Se calcula la varianza para cada caso, por ejemplo, la varianza
     de cada fila
268     var.of.reestimates<-apply(jackknife.ests, 1, var)
269     ## Se re-escala
270     jackknife.var<-((n-1)^2/n)* var.of.reestimates
271     ## Se aplica sqrt() para obtener el error estándar
272     jackknife.stderrs<-sqrt(jackknife.var)
273     return(jackknife.stderrs)
274 }
275
```

```
276 ##############################################################################
277 ######## coeficientes beta, (mínimos cuadrados ponderados) ###########
278 ##############################################################################
279 bj<-est.coefs(datos)
280
281 ##############################################################################
282 ################## Se calculan los errores estándar ######################
283 ##############################################################################
284 ee.jack<-jackknife(estimator=est.coefs, data=datos)
285
286
287 ##############################################################################
288 ###### Cálculo del error estándar mediante el método bootstrap ######
289 ##############################################################################
290
291 ##############################################################################
292 ###### Función para obtener los coeficientes mediante mínimos ########
293 ###### cuadrados ponderados ##################################################
294 ##############################################################################
295 bc<-function(formula, data, indices) {
296     muesboot<-data[indices, ] # permite seleccionar la muestra
    bootstrap
297     fit<-lm(formula, data=muesboot, weight=CW)
298     return(coef(fit))
299 }
300
301 ##############################################################################
302 ######## Se realiza el bootstrapping con la longitud R=10000 #########
303 ##############################################################################
304 results<-boot(data=datos, statistic=bc, stype=c("i"),
305                 R=10000, formula=log(Failuretime) ~ -1 + DuctileCI +
    GrayCI + Polyethylene + Sidewalk + Normal + Length + Diameter)
306
307 ##############################################################################
308 ######### coeficientes beta, (mínimos cuadrados ponderados) ##########
309 ##############################################################################
310 bt<-as.data.frame(results$t0)
311
312 ##############################################################################
313 ################## Se calculan los errores estándar ######################
314 ##############################################################################
315 ee.boot<-as.data.frame(apply(results$t, 2, sd))
316
317 ##############################################################################
318 ############ Construcción de los intervalos de confianza #################
319 ##############################################################################
320
```

```
321 ##############################################################################
322 #### Se decide construir los intervalos de confianza a través de ####
323 #### errores estándar de tipo Bootstrap, aunque se podrían construir#
324 #### mediante los errores estándar de tipo Jackknife ################
325 ##############################################################################
326 level<-0.95
327 probinf<-(1-level)/2
328 probsup<- 1-probinf
329 z<-qnorm(c(probinf, probsup))
330 int<-cbind(bt + z[1]*ee.boot, bt, bt + z[2]*ee.boot )
331 print(int)
332
333 ##############################################################################
334 #### O directamente mediante las funciones del package bootstrap ####
335 ##############################################################################
336 boot.ci(results, type="norm", conf=0.95, index=1)
337 boot.ci(results, type="norm", conf=0.95, index=2)
338 boot.ci(results, type="norm", conf=0.95, index=3)
339 boot.ci(results, type="norm", conf=0.95, index=4)
340 boot.ci(results, type="norm", conf=0.95, index=5)
341 boot.ci(results, type="norm", conf=0.95, index=6)
342 boot.ci(results, type="norm", conf=0.95, index=7)
343
344
345 ##############################################################################
346 ##############################################################################
347 ############## ESTIMADOR CONSTANTE LOCAL Y LINEAL LOCAL ##################
348 ##############################################################################
349 ##############################################################################
350
351 ##############################################################################
352 ######## Se transforman los datos a escala de tiempos base ###########
353 ##############################################################################
354 mdatx<-cbind(datos[ ,14], datos[ ,15], datos[ ,16], datos[ ,17], datos
      [ ,18], datos[ ,1], datos[ ,2])
355
356 ##############################################################################
357 ################## Se traspone la matriz de covariables ##################
358 ##############################################################################
359 mdatx<-as.matrix(mdatx)
360 mdatx<-t(mdatx)
361
362 ##############################################################################
363 ######## Se trasponen los betas y se transformarman en matriz ########
364 ##############################################################################
365 bt<-as.matrix(bt)
366 bt<-t(bt)
```

```
367
368 ##############################################################################
369 ################## Función exponencial y su traspuesta ######################
370 ##############################################################################
371 fet<-bt%*%mdatx
372 fet<-t(fet)
373
374 ##############################################################################
375 ############ Tiempos de fallo de la base de datos original ###################
376 ##############################################################################
377 Ti<-as.matrix(datos[ ,4])
378 Ti<-log(Ti)### se le aplica logaritmos naturales
379
380 ##############################################################################
381 ############## Tiempos de fallo en escala logarítmica base ###################
382 ##############################################################################
383 lnT0<-fet+Ti
384 T0<-exp(lnT0)### tiempo de fallo en escala base
385
386 ##############################################################################
387 ######## Tiempos de truncamiento de la base de datos original ########
388 ##############################################################################
389 Li<-as.matrix(datos[ ,13])
390 Li<-log(Li)### se le aplica logaritmos naturales
391
392 ##############################################################################
393 ######## Tiempo de truncamiento en escala logarítmica base  ##########
394 ##############################################################################
395 lnL0<-fet+Li
396 L0<-exp(lnL0)### tiempo de truncamiento en escala base
397
398 ##############################################################################
399 ########################## Nueva base de datos ###############################
400 ##############################################################################
401 cens<-as.matrix(datos[ ,11])### código de censura
402 tuplanueva<-cbind(T0,cens,L0)### nueva base de datos
403 tuplanueva<-as.data.frame(tuplanueva)### se transforma a dataframe
404
405 ##############################################################################
406 ########## Se guarda la nueva base de datos en un fichero ############
407 ##############################################################################
408 write.csv(tuplanueva, file="tuplanueva.csv",quote=FALSE, row.names=
    FALSE)
409
410
411 ##############################################################################
412 ############## Se procede al cálculo de dicho estimador ################
```

```
413 ##########################################################################################
414 datos2<-csv.get('tuplanueva.csv')
415
416 ##########################################################################################
417 ################## Se pasan los datos en forma de matriz ##################
418 ##########################################################################################
419 datos2<-as.matrix(datos2)
420
421 ##########################################################################################
422 ###################### Se ordena de forma creciente ######################
423 ##########################################################################################
424 datos2<-datos[order(datos2[ ,1]), ]
425 nombres<-c('Ti','deltai','Li')
426 colnames(datos2)<-nombres
427 T0<-datos2[ ,1] ##tiempo base de fallo o censura
428 delta0<-datos2[ ,2] ##código de censura, 1 o 0
429 L0<-datos2[ ,3] ##tiempo base de truncamiento
430
431 ##########################################################################################
432 ############################## Proceso de riesgo ##############################
433 ##########################################################################################
434 riesgo<-function(t,Li,Ti){
435     # calcula el tamaño del conjunto de riesgo en cada coordenada de t
    .
436     # Li=truncamiento
437     # Ti=fallo o censura
438     # t es un vector o un escalar
439     Y.n<-sapply(1:length(t),function(i) {length(which(Li<t[i] & t[i]<=
    Ti))})
440     Y.n
441 }
442
443 ##########################################################################################
444 ##################### Función de supervivencia empírica ##################
445 ##########################################################################################
446 S0<-function(t, Li, deltai, Ti){
447     # se calcula la función de riesgo h(i)=nºfallos en x1(i)/
    nºindividuos en riesgo en x1(i)
448     # t es un vector o un escalar
449     ri<-riesgo(Ti, Li, Ti)
450     hi<-deltai/ri # función de riesgo evaluado en el vector de tiempos
     de fallo
451     H<-cumsum(hi) # función de riesgo acumulado evaluado en el vector
    de tiempos de fallo
452     S0.i<-exp(-H) # función de supervivencia evaluada en el vector de
    tiempos de fallo
453     S0.t<-approx(x=Ti, y=S0.i, xout=t, rule=2)
```

```
    S0.t$y  # el valor de la funcion de supervivencia en t
}

#############################################################################
############## Integración numérica, Método de Simpson ##################
#############################################################################
simpson<-function(fun, a, b, n=100) {
    # integración numérica utilizando la regla de Simpson
    # se supone que a < b y que n es un entero positivo
    h<-(b-a)/n
    x <- seq(a, b, by=h)
    if (n==2) {
        s<-fun(x[1]) + 4*fun(x[2]) + fun(x[3])
    } else {
        s<-fun(x[1]) + fun(x[n+1]) + 2*sum(fun(x[seq(2, n, by=2)])) +
    4*sum(fun(x[seq(3, n-1, by=2)]))
    }
    s<-s*h/3
    return(s)
}

#############################################################################
############## Función tipo núcleo de Epanechnikov ######################
#############################################################################
epanech<-function(u){0.75*(1 - u^2)*(abs(u)<1)}

#############################################################################
################## Función del Estimador Constante Local #################
#############################################################################
S.LC<-function(t, Li, deltai, Ti, b){
    #t es un vector
    num.i<-function(x){
        #para cada escalar devuelve la integral del numerador
        int.num<-function(s) {return(epanech((x-s)/b)*S0(s, Li, deltai
    , Ti)*riesgo(s, Li, Ti))}
        return(simpson(int.num, 0, max(Ti)))
    }
    numerador<-sapply(1:length(t), function(i) {num.i(t[i])})
    den.i<-function(x){
        #para cada escalar devuelve la integral del denominador
        int.den<-function(s) {return(epanech((x-s)/b)*riesgo(s,Li,Ti))
    }
        return(simpson(int.den, 0, max(Ti)))
    }
    denominador<-sapply(1:length(t), function(i) {den.i(t[i])})
    out<-numerador/denominador
    return(out)
```

```
}


#############################################################################
################## Función del Estimador Lineal Local ######################
#############################################################################
S.LL<-function(t, Li, deltai, Ti, b){
    #t es un vector
    a0.i<-function(x){
        #para cada escalar devuelve la integral
        int.a0<-function(s) {return(epanech((x-s)/b)*riesgo(s, Li, Ti)
    )}
        return(simpson(int.a0, 0, max(Ti)))
    }
    a0<-sapply(1:length(t), function(i) {a0.i(t[i])})

    a1.i<-function(x){
        #para cada escalar devuelve la integral
        int.a1<-function(s) {return(epanech((x-s)/b)*(x-s)*riesgo(s,
    Li, Ti))}
        return(simpson(int.a1, 0, max(Ti)))
    }
    a1<-sapply(1:length(t), function(i) {a1.i(t[i])})

    a2.i<-function(x){
        #para cada escalar devuelve la integral
        int.a2<-function(s) {return(epanech((x-s)/b)*(x-s)*(x-s)*
    riesgo(s, Li, Ti))}
        return(simpson(int.a2, 0, max(Ti)))
    }
    a2<-sapply(1:length(t), function(i) {a2.i(t[i])})

    G0.i<-function(x){
        #para cada escalar devuelve la integral
        int.G0<-function(s) {return(epanech((x-s)/b)*S0(s, Li, deltai,
     Ti)*riesgo(s, Li, Ti))}
        return(simpson(int.G0, 0, max(Ti)))
    }
    G0<-sapply(1:length(t), function(i) {G0.i(t[i])})

    G1.i<-function(x){
        #para cada escalar devuelve la integral
        int.G1<-function(s) {return(epanech((x-s)/b)*S0(s, Li, deltai,
     Ti)*(x-s)*riesgo(s, Li, Ti))}
        return(simpson(int.G1, 0, max(Ti)))
    }
    G1<-sapply(1:length(t), function(i) {G1.i(t[i])})
```

```
540
541     out<-(a0*G0-a1*G1)/(a0*a2-a1*a1)
542     return(out)
543 }
544
545
546 ##############################################################################
547 ########## Criterio de selección del parámetro de suavizado ##########
548 ##############################################################################
549 b1<-sd(T0)*(4/length(T0))^(1/3)
550
551 ##############################################################################
552 ############### Se calcula el estimador en un intervalo ################
553 ##############################################################################
554 t<-seq(0, max(T0), length = 100)
555
556 ##############################################################################
557 ###### Funciones de supervivencia empírica y suavizadas según ########
558 ###### los valores de t y el parámetro de suavizado considerado ######
559 ##############################################################################
560 s.t0<-S0(t, L0, delta0, T0) # empírica
561 s.t1<-sapply(1:100,function(i) S.LL(t[i], L0, delta0, T0, b1)) #
    suavizada
562
563 ##############################################################################
564 #### Gráfica de la función de supervivencia empírica y suavizada ####
565 ##############################################################################
566 win.graph()
567 plot(s.t0, ylim=c(0,1), type="l", pch=1, xaxt="n", yaxt="n", axes=T,
    ann=T, xlab="Tiempos",
568       ylab="Función de supervivencia empírica y suavizada", cex.lab
    =0.6, cex.main=1.0, lwd=2)
569 axis(1, cex.axis=0.6)
570 axis(2, cex.axis=0.6)
571 lines(s.t1, type="l", lty=2, pch=2, lwd=2)
572
573 ##############################################################################
574 ##############################################################################
575 ########################## ESTUDIO DE SIMULACIÓN ############################
576 ##############################################################################
577 ##############################################################################
578
579 nl<-50 ## ó 100 ó 200 tamaño muestral
580 cel<-0.9 ### variará según el tamaño muestral y el nivel de censura
581 hel<-0.4 ### variará según el tamaño muestral y el nivel de censura
582 err1<-log(rweibull(n, shape=1, scale=1)) ### para el caso exponencial
583 err2<-log(rweibull(n, shape=0.5, scale=1)) ### para weibull con
```

```
    parámetro de forma igual a 0,5.
err3<-log(rweibull(n, shape=5, scale=1)) ### para weibull con
    parámetro de forma igual a 5.
beta1<-1
beta2<-3

#############################################################################
########## Se simulan todas las muestras y el ajuste de Cox #########
#############################################################################
simula<-function(n){
    ### Se generan los datos del modelo de Cox ###
    x1<-runif(n, min=0, max=2)
    x2<-runif(n, min=3, max=9)
    ### error min valor extremo ###
    err1<-log(rweibull(n, shape=1, scale=1))
    ### Tiempo del evento con distribución de errores de tipo min
    extreme value ###
    time<-1*x1 + 3*x2 + err
    ### Se rellena con 1 si se observa el evento ###
    censoring<-rbinom(n, 1, cel)
    ### Se crean posibles tiempos de truncamiento ###
    truncated<-runif(n, min=0, max=(quantile(time, hel)))
    ### Se guarda todo en una tabla ###
    datos<-cbind(x1, x2, truncated, time, censoring)
    ### El tiempo de truncamiento es menor que el censurado ###
    datosphr<-as.data.frame(datos[datos[ ,4] > datos[ ,3], ])
    ### Se muestra el número de fallos ###
    cen<-as.matrix(table(datosphr$censoring))
    list(cen=cen, datosphr=datosphr)
}

#############################################################################
##### Se generan bases de datos aleatoriamente cumpliendo con el ####
##### porcentaje de censura deseado ##################################
#############################################################################
datosfiltrados<-function(M){
    resul<-list()
    datdef<-list()
    for (i in 1:M){
        resul[[i]]<-simula(nl)
        datdef[[i]]<-result[[i]]$datosphr
    }
    return(datdef)
}

M<-2000
datos3<-datosfiltrados(M)
```

```
628
629 ##############################################################################
630 #### Se ordenan las bases de datos en tiempos de fallo creciente ####
631 ##############################################################################
632 orden<-list() ## se guardan en forma de lista
633 for (i in 1:M){
634     orden[[i]]<-order(datos3[[i]]$time)
635     datos3[[i]]<-datos3[[i]][orden[[i]], ]
636 }
637
638 ##############################################################################
639 ######## Se guardan las bases de datos en un fichero .Rdata #########
640 ##############################################################################
641 save(datos3, file="censura_size_param.Rdata")
642
643
644 ##############################################################################
645 ######## CÁLCULO DEL ASE PARA EL MODELO SEMI-PARAMÉTRICO AFT #########
646 ##############################################################################
647
648 ##############################################################################
649 ########## Se carga la base de datos de matriz por bloques ##########
650 ##############################################################################
651 datos4<-local(get(load('censura_size_param.RData')))
652 datos4<-as.matrix(datos4)
653
654 ##############################################################################
655 ########## Se implementa la función de riesgo acumulado  ############
656 ##############################################################################
657 tf<-list()
658 cc<-list()
659 tc<-list()
660 nsiz<-list()
661 for (i in 1:M){
662     tf[[i]]<-exp(datos4[[i]][ ,4]) ##tiempo de fallo
663     cc[[i]]<-datos4[[i]][ ,5] ##código de censura
664     tc[[i]]<-exp(datos4[[i]][ ,3]) ##tiempo de truncamiento
665     nsiz[[i]]<-nrow(datos4[[i]]) ## tamaño de cada muestra
666 }
667
668 hazcum2<-function(tf,cc,tc,nsiz){
669     risk<-function(x,tc,tf){sapply(1:length(x), function(i) length(
    which(tc<x[i] & x[i]<=tf)))}
670     r<-risk(x=tf, tc=tc, tf=tf)  # tamaño del conjunto de riesgo para
    cada elemento
671     Hi<-cumsum(cc/r) # este es el estimador de Nelson-Aalen de la
    función de riesgo acumulado
```

```
672     return(Hi)
673 }
674
675 Ha2<-list()
676 for (i in 1:M){
677     Ha2[[i]]<-hazcum2(tf[[i]], cc[[i]], tc[[i]], nsiz[[i]])
678 }
679
680 ##############################################################################
681 ###################### Función de distribución ###############################
682 ##############################################################################
683 Fdist2<-list()
684 for (i in 1:M){
685     Fdist2[[i]]<-(1-exp(-Ha2[[i]]))
686 }
687
688 ##############################################################################
689 ######################### Se calculan los pesos ##############################
690 ##############################################################################
691 np<-list()
692 Wp<-list()
693 datospe<-list()
694 for (i in 1:M){
695     np[[i]]<-length(Fdist2[[i]])
696     Wp[[i]]<-rep(0, np[[i]])
697 }
698
699 for (i in 1:M){
700     Wp[[i]][1]<-Fdist2[[i]][1]
701     for (j in 2:np[[i]]){
702         Wp[[i]][j]<-Fdist2[[i]][j] - Fdist2[[i]][j-1]
703     }
704     datospe[[i]]<-cbind(datos4[[i]], Wp[[i]])
705 }
706 Wlist<-as.matrix(Wp)
707 X<-list()
708 Z<-list()
709 for (i in 1:M){
710     X[[i]]<-cbind(datospe[[i]][1],datospe[[i]][2])
711     Z[[i]]<-exp(datospe[[i]][4])
712 }
713
714 Xv<-as.matrix(X)
715 Zv<-as.matrix(Z)
716
717 ##############################################################################
718 #### Esta es la función que tiene forma del estimador de mínimos ####
```

```
#### cuadrados ponderados ##################################################
############################################################################
est.coefs2<-function(x,y,w) {
    return(lsfit(x, y, wt=w, intercept=FALSE, yname=NULL)$coef)
}
coeff<-list()
for (i in 1:M){
    coeff[[i]]<-est.coefs2(Xv[[i]], log(Zv[[i]]), Wlist[[i]])
}

############################################################################
###### Se calcula la media de todos los coeficientes calculados ############
############################################################################
sumest1<-matrix(0,M,0)
sumest2<-matrix(0,M,0)
for (i in 1:M){
    sumest1[i]<-sum(as.matrix(coeff[[i]][1]))
    sumest2[i]<-sum(as.matrix(coeff[[i]][2]))
}
meanest1<-(1/M)*sum(sumest1)
meanest2<-(1/M)*sum(sumest2)

############################################################################
########### Se calcula el sesgo y la desviación estándar ###################
############################################################################
biasest1<-meanest1-1
biasest2<-meanest2-3
sumsd1<-matrix(0,M,0)
sumsd2<-matrix(0,M,0)
for (i in 1:M){
    sumsd1[i]<-sum((as.matrix(coeff[[i]][1]) - meanest1)^2)
    sumsd2[i]<-sum((as.matrix(coeff[[i]][2]) - meanest2)^2)
}
sdest1<-sqrt((1/(M-1))*sum(sumsd1))
sdest2<-sqrt((1/(M-1))*sum(sumsd2))

############################################################################
##################### El error cuadrático medio ############################
############################################################################
mse1<-(sdest1)^2 + (biasest1)^2
mse2<-(sdest2)^2 + (biasest2)^2

############################################################################
############ Coeficientes para el ASE y para el box-plot ###################
############################################################################
coeff<-as.matrix(coeff)
coef_censura_size_semi<-as.data.frame(coeff)
```

```
766
767 ##############################################################################
768 ######## Se guardan las bases de datos en un fichero .Rdata #########
769 ##############################################################################
770 save(coef_censura_size_semi, file="coef_censura_size_semi.Rdata")
771
772 ##############################################################################
773 ##### Se carga la base de datos de matriz por bloques en forma ######
774 ##############################################################################
775 datos5<-local(get(load('censura_size_param.RData')))
776 datos5<-as.matrix(datos5)
777 coef1<-local(get(load('coef_censura_size_semi.Rdata')))
778 coef1<-as.matrix(coef1)
779
780 ##############################################################################
781 ########## Se transforman los datos a tiempos base T0 y L0 ##########
782 ##############################################################################
783 mdatx2<-list()
784 bt2<-list()
785 fet2<-list()
786 Ti2<-list()
787 T02<-list()
788 Li2<-list()
789 L02<-list()
790 cens2<-list()
791 datosnew<-list()
792 for (i in 1:M){
793     mdatx2[[i]]<-t(cbind(datos5[[i]][ ,1],datos5[[i]][ ,2])) ##matriz
    de covariables
794     bt2[[i]]<-t(coef1[[i]]) ##matriz de coeficientes
795     fet2[[i]]<-t(-bt2[[i]]%%mdatx2[[i]]) ##función exponencial
796     Ti2[[i]]<-exp(datos5[[i]][ ,4]) ##tiempos de fallo
797     T02[[i]]<-Ti2[[i]]*exp(fet2[[i]]) ##tiempos de fallo base
798     Li2[[i]]<-exp(datos5[[i]][ ,3]) ##tiempos de truncamiento
799     L02[[i]]<-Li2[[i]]*exp(fet2[[i]]) ##tiempos de truncamiento base
800     cens2[[i]]<-datos5[[i]][ ,5] ##código de censura
801     datosnew[[i]]<-cbind(T02[[i]], cens2[[i]], L02[[i]]) ##datos
    finales a escala base
802 }
803
804 ##############################################################################
805 ### Se ordenan las bases de datos en tiempos de fallo crecientes ####
806 ##############################################################################
807 orden2<-list()
808 for (i in 1:M){
809     orden2[[i]]<-order(datosnew[[i]][ ,1])
810     datosnew[[i]]<-datosnew[[i]][orden2[[i]], ]
```

```
    nombres<-c('Ti','deltai','Li')
    colnames(datosnew[[i]])<-nombres
}

#############################################################################
############# Parámetros de suavizado para todos los casos ############
#############################################################################
b1l<-list()
for (i in 1:M){
    b1l[[i]]<-sd(datosnew[[i]][ ,1])*(4/length(datosnew[[i]][ ,1]))
    ^(1/3)
}

#############################################################################
### Se calcula el estimador lineal local en los tiempos de fallo ####
#############################################################################
s.t2<-list()
for (i in 1:M){
    s.t2[[i]]<-S.LL(datosnew[[i]][ ,1], datosnew[[i]][ ,3], datosnew[[
    i]][ ,2], datosnew[[i]][ ,1], b1l[[i]])
}

#############################################################################
################## Ahora con los coeficientes reales ###################
#############################################################################
coef1r<-as.matrix(coef1)
for (i in 1:M){
    coef1r[[i]][1]<-1
    coef1r[[i]][2]<-3
}
mdatxr<-list()
btr<-list()
fetr<-list()
Tir<-list()
T0r<-list()
Lir<-list()
L0r<-list()
censr<-list()
datosr<-list()
for (i in 1:M){
    mdatxr[[i]]<-t(cbind(datos5[[i]][ ,1],datos5[[i]][ ,2])) ##matriz
    de covariables
    btr[[i]]<-t(coef1r[[i]]) ##matriz de coeficientes reales
    fetr[[i]]<-t(-btr[[i]]%%mdatxr[[i]]) ##función exponencial
    Tir[[i]]<-exp(datos5[[i]][ ,4]) ##tiempos de fallo
    T0r[[i]]<-Tir[[i]]*exp(fetr[[i]]) ##tiempos de fallo base
    Lir[[i]]<-exp(datos5[[i]][,3]) ##tiempos de truncamiento
```

```
    L0r[[i]]<-Lir[[i]]*exp(fetr[[i]]) ##tiempos de truncamiento base
    censr[[i]]<-datos5[[i]][,5] ##código de censura
    datosr[[i]]<-cbind(T0r[[i]], censr[[i]], L0r[[i]]) ##datos finales
    a escala base
}

#############################################################################
### Se ordenan las bases de datos en tiempos de fallo crecientes ####
#############################################################################
orden3<-list()
for (i in 1:M){
    orden3[[i]]<-order(datosr[[i]][ ,1])
    datosr[[i]]<-datosr[[i]][orden3[[i]], ]
}

#############################################################################
########## Se calcula la función de supervivencia teórica ###########
#############################################################################
s.t1r<-list()
for (i in 1:M){
    s.t1r[[i]]<-exp(-(datosr[[i]][ ,1]))
}

#############################################################################
################## ASE para el modelo semi-paramétrico ################
#############################################################################
sumat1<-matrix(0,M,0)
for (i in 1:M){
    sumat1[[i]]<-(1/length(s.t2[[i]]))*sum((s.t2[[i]] - s.t1r[[i]])^2)
}

#############################################################################
######### Se guardan todos los ASE para crear los boxplot ############
#############################################################################
ASE_censura_size_semi<-as.data.frame(sumat1)
######### se guardan las bases de datos en un fichero .Rdata ########
save(ASE_censura_size_semi, file = "ASE_censura_size_semi.Rdata")


#############################################################################
############ CÁLCULO DEL MSE PARA EL MODELO PARAMÉTRICO AFT ###########
#############################################################################

########## Se carga la base de datos de matriz por bloques ##########
datos6 <- local(get(load('censura_size_param.RData')))
datos6 <- as.matrix(datos6)

```

```
#############################################################################
######## Se ajusta el modelo AFT con distribución exponencial #########
#############################################################################
funexp<-function(datos){
    aftmexp<-list()
    res<-list()
    M<-2000
    for (i in 1:M){
        aftmexp[[i]]<-weibreg(Surv(exp(truncated), exp(time),
    censoring) ~ x1 + x2, data=datos[[i]], shape=1)
        summary(aftmexp[[i]])
        res[[i]]<-coef(aftmexp[[i]])
    }
    return(res)
}

#############################################################################
################ Se pasan los resultados a una matriz ###################
#############################################################################
resexp<-as.matrix(funexp(datos6))

#############################################################################
##### Se calcula la media de todos los coeficientes calculados ######
#############################################################################
sumest1exp<-matrix(0,M,0)
sumest2exp<-matrix(0,M,0)
for (i in 1:M){
    sumest1exp[i]<-sum(resexp[[i]][1])
    sumest2exp[i]<-sum(resexp[[i]][2])
}
meanest1exp<-abs((1/M)*sum(sumest1exp))
meanest2exp<-abs((1/M)*sum(sumest2exp))

#############################################################################
######################## Se calcula el sesgo ################################
#############################################################################
biasest1exp<-meanest1exp-1
biasest2exp<-meanest2exp-3

#############################################################################
################## Se calcula la desviación estándar ####################
#############################################################################
sumsd1exp<-matrix(0,M,0)
sumsd2exp<-matrix(0,M,0)
for (i in 1:M){
    sumsd1exp[i]<-sum((resexp[[i]][1] + meanest1exp)^2)
    sumsd2exp[i]<-sum((resexp[[i]][2] + meanest2exp)^2)
```

```
947 }
948 sdest1exp<-sqrt((1/(M-1))*sum(sumsd1exp))
949 sdest2exp<-sqrt((1/(M-1))*sum(sumsd2exp))
950
951 ##############################################################################
952 ################## Se calcula el error cuadrático medio ######################
953 ##############################################################################
954 mse1exp<-(sdest1exp)^2 + (biasest1exp)^2
955 mse2exp<-(sdest2exp)^2 + (biasest2exp)^2
956
957
958 ##############################################################################
959 ################## CÁLCULO DEL ASE PARA EL MODELO PH #########################
960 ##############################################################################
961
962 ########## Se carga la base de datos de matriz por bloques ###########
963 datos7<-local(get(load('censura_size_param.RData')))
964 datos7<-as.matrix(datos7)
965
966 ##############################################################################
967 ###### Se ajusta el modelo de Cox para las 2000 bases de datos #######
968 ##############################################################################
969 funCox<-function(datos){
970     survobj<-list()
971     res<-list()
972     M<-2000
973     for (i in 1:M){
974         survobj[[i]]<-coxph(Surv(exp(truncated),exp(time),censoring) ~
      x1 + x2, data=datos[[i]], method="breslow")
975         summary(survobj[[i]])
976         res[[i]]<-summary(survobj[[i]])$coeff
977     }
978     return(res)
979 }
980
981 ##############################################################################
982 ################## Se pasa el resultado a una matriz #########################
983 ##############################################################################
984 rescox<-as.matrix(funCox(datos7))
985
986 ##############################################################################
987 #### Se calcula la media de todos los coeficientes calculados ######
988 ##############################################################################
989 sumest1cox<-matrix(0,M,0)
990 sumest2cox<-matrix(0,M,0)
991 for (i in 1:M){
992     sumest1cox[i]<-sum(rescox[[i]][1])
```

```
    sumest2cox[i]<-sum(rescox[[i]][2])
}
meanest1cox<-(1/M)*sum(sumest1cox)
meanest2cox<-(1/M)*sum(sumest2cox)

#############################################################################
########################## Se calcula el sesgo ##############################
#############################################################################
biasest1cox<-meanest1cox + 1
biasest2cox<-meanest2cox + 3

#############################################################################
################## Se calcula la desviación estándar ########################
#############################################################################
sumsd1cox<-matrix(0,M,0)
sumsd2cox<-matrix(0,M,0)
for (i in 1:M){
    sumsd1cox[i]<-sum((rescox[[i]][1] - meanest1cox)^2)
    sumsd2cox[i]<-sum((rescox[[i]][2] - meanest2cox)^2)
}
sdest1cox<-sqrt((1/(M-1))*sum(sumsd1cox))
sdest2cox<-sqrt((1/(M-1))*sum(sumsd2cox))

#############################################################################
################## Se calcula el error cuadrático medio #####################
#############################################################################
mse1cox<-(sdest1cox)^2 + (biasest1cox)^2
mse2cox<-(sdest2cox)^2 + (biasest2cox)^2

#############################################################################
##### Se guardan todos los coeficientes para creae los boxplot ######
#############################################################################
coef_censura_size_cox<-as.data.frame(rescox)

#############################################################################
######## Se guardan las bases de datos en un fichero .Rdata ##########
#############################################################################
save(coef_censura_size_cox, file="coef_censura_size_cox.Rdata")

#############################################################################
######## Se carga la base de datos de matriz por bloques #############
#############################################################################
datos8<-local(get(load('censura_size_param.RData')))
datos8<-as.matrix(datos8)
coef1cox<-local(get(load('coef_censura_size_cox.Rdata')))
coef1cox<-as.matrix(coef1cox)

```

```
1040 ##########################################################################
1041 ########## Se calculan las 2000 curvas base estimadas y real ##########
1042 ##########################################################################
1043 estcox<-function(datos){
1044     fit1<-list()
1045     So1<-list()
1046     M<-2000
1047     for (i in 1:M){
1048         fit1[[i]]<-coxph(Surv(exp(truncated), exp(time), censoring) ~
     x1 + x2, data=datos[[i]], method="breslow")
1049         So1[[i]]<-exp(-basehaz(fit1[[i]])$hazard)
1050     }
1051     return(So1)
1052 }
1053
1054 ##########################################################################
1055 ################ Ahora con los coeficientes reales ######################
1056 ##########################################################################
1057 coef1rcox<-as.matrix(coef1cox)
1058 for (i in 1:M){
1059     coef1rcox[[i]][1]<- 1
1060     coef1rcox[[i]][2]<- 3
1061 }
1062 mdatxrcox<-list()
1063 btrcox<-list()
1064 fetrcox<-list()
1065 Tircox<-list()
1066 T0rcox<-list()
1067 Lircox<-list()
1068 L0rcox<-list()
1069 censrcox<-list()
1070 datosrcox<-list()
1071 for (i in 1:M){
1072     mdatxrcox[[i]]<-t(cbind(datos8[[i]][ ,1], datos8[[i]][ ,2])) ##
     matriz de covariables
1073     btrcox[[i]]<-t(coef1rcox[[i]]) ##matriz de coeficientes reales
1074     fetrcox[[i]]<-t(-btrcox[[i]]%%mdatxrcox[[i]]) ##función
     exponencial
1075     Tircox[[i]]<-exp(datos8[[i]][ ,4]) ##tiempos de fallo
1076     T0rcox[[i]]<-Tircox[[i]]*exp(fetrcox[[i]]) ##tiempos de fallo base
1077     Lircox[[i]]<-exp(datos8[[i]][ ,3]) ##tiempos de truncamiento
1078     L0rcox[[i]]<-Lircox[[i]]*exp(fetrcox[[i]]) ##tiempos de
     truncamiento base
1079     censrcox[[i]]<-datos8[[i]][ ,5] ##código de censura
1080     datosrcox[[i]]<-cbind(T0rcox[[i]], censrcox[[i]], L0rcox[[i]]) ##
     datos finales a escala base
1081 }
```

```
1082
1083 ##############################################################################
1084 ### Se ordenan las bases de datos en tiempos de fallo crecientes ####
1085 ##############################################################################
1086 orden4<-list()
1087 for (i in 1:M){
1088     orden4[[i]]<-order(datosrcox[[i]][,1])
1089     datosrcox[[i]]<-datosrcox[[i]][orden4[[i]], ]
1090 }
1091
1092 ##############################################################################
1093 ########## Se calcula la función de supervivencia teórica ############
1094 ##############################################################################
1095 S2<-list()
1096 for (i in 1:M){
1097     S2[[i]]<-exp(-(datosrcox[[i]][,1]))
1098 }
1099
1100 ##############################################################################
1101 ############### Curva de supervivencia base estimada ####################
1102 ##############################################################################
1103 S1<-estcox(datos8)
1104
1105 ##############################################################################
1106 ########### ASE para el modelo semi-paramétrico de Cox ###############
1107 ##############################################################################
1108 sumat1cox<-matrix(0,M,0)
1109 M<-2000
1110 for (i in 1:M){
1111     sumat1cox[[i]]<-(1/length(S1[[i]]))*sum((S1[[i]] - S2[[i]])^2)
1112 }
1113
1114 ##############################################################################
1115 ##### Se guardan todos los ASE y la curva de supervivencia para #####
1116 ##### los box-plots ##########################################################
1117 ASE_censura_size_cox <- as.data.frame(sumat1cox)
1118
1119 ##############################################################################
1120 ######## Se guardan las bases de datos en un fichero .Rdata #########
1121 ##############################################################################
1122 save(ASE_censura_size_cox, file="ASE_censura_size_cox.Rdata")
1123
1124 ##############################################################################
1125 ############# Box-plots para el modelo de cox y el AFT ################
1126 ######### Se cargan los coeficientes de ambas estimaciones ###########
1127 ##############################################################################
1128 coxbp<-local(get(load('ASE_censura_size_cox.Rdata')))
```

```
semipbp<-local(get(load('ASE_censura_size_semi.Rdata')))
coxbp<-as.matrix(coxbp)
semipbp<-as.matrix(semipbp)

#############################################################################
########### Se pasan a dataframe y se le ponen los nombres ##############
#############################################################################
datosbp<-data.frame(coxbp, semipbp)
nombres<-c('PH','AFT')
colnames(datosbp)<-nombres

#############################################################################
############################ Diagramas box-plots ############################
#############################################################################
boxplot(datosbp,xlab="n=size and censoring level censoring %",ylab="
    ASE")
```

Apéndice B

Código en R del estudio de simulación del Capítulo 3

```
##############################################################################
##############################################################################
################### IMPLEMENTACIÓN DEL CAPÍTULO 3 ############################
##############################################################################
##############################################################################

##############################################################################
############################## Libros necesarios #############################
##############################################################################
library(foreign)
library(Matrix)
library(gplots)
library(lattice)
library(itertools)

##############################################################################
### Función que realiza la inversa de la primitiva de la función ####
### bañera ###################################################################
##############################################################################
Inv.Lb<-function(s){
    a<-5
    if (s<a^3/3) {t<-a-(a^3-3*s)^(1/3)}
    if (s==a^3/3) {t<-a}
    if (s>a^3/3) {t<-a+(3*s-a^3)^(1/3)}
    return(t)
}

##############################################################################
### Función que realiza la inversa de la primitiva de la función ####
```

```
### dos cambios ##############################################################
Lambdadc<-function(t) return((1/12)*t^(4)-(4/3)*t^(3)+6*t^(2)+t)
Inv.Ldc<-function(u){
    L.u<-function(t) return(Lambdadc(t)-u)
    t.u<-uniroot(L.u,c(0,1e+10))$root
    return(t.u)
}

##############################################################################
### Función que realiza la inversa de la primitiva de la función ####
### periódica tipo I #########################################################
##############################################################################
Lambdap<-function(t) return(100*t+(100/pi)*sin(0.5*pi*t))
Inv.Lp<-function(u){
    L.u<-function(t) {return(Lambdap(t)-u)}
    t.u<-uniroot(L.u,c(0,1e+10))$root
    return(t.u)
}

##############################################################################
### Función que realiza la inversa de la primitiva de la función ####
### periódica tipo II ########################################################
##############################################################################
Lambdap2<-function(t) return(100*t+(100/(3*pi))*sin(1.5*pi*t))
Inv.Lp2<-function(u){
    L.u<-function(t) {return(Lambdap2(t)-u)}
    t.u<-uniroot(L.u,c(0,1e+10))$root
    return(t.u)
}

##############################################################################
############ Estimador rocof con núcleo de Epanechnikov ##############
##############################################################################
Lepanechnikov<-function(t,datos,h){
    epa<-function(v){return(0.75*(1-v^2)*(abs(v)<=1))}
    v<-(t-datos)/h
    Nepa<-epa((t-datos)/h)
    lambdaestima<-(1/h)*sum(Nepa)
    lambdaestima
}

##############################################################################
###### Derivada del estimador rocof con núcleo de Epanechnikov ######
##############################################################################
dLepanechnikov<-function(t,datos,h){
    depa<-function(v){return((-1.5)*v*(abs(v)<=1))}
    v<-(t-datos)/h
```

```
77      dNepa<-depa((t-datos)/h)
78      dlambdaestima<-(1/(h^2))*sum(dNepa)
79      dlambdaestima
80  }
81
82  ##############################################################################
83  ################## Estimador rocof con núcleo Biweight ######################
84  ##############################################################################
85  Lbiweight<-function(t,datos,h){
86      biw<-function(v){return(0.9375*((1-v^2)^2)*(abs(v)<=1))}
87      v<-(t-datos)/h
88      Nbiw<-biw((t-datos)/h)
89      lambdaestima<-(1/h)*sum(Nbiw)
90      lambdaestima
91  }
92
93  ##############################################################################
94  ######### Derivada del estimador rocof con núcleo Biweight ###########
95  ##############################################################################
96  dLbiweight<-function(t,datos,h){
97      dbiw<-function(v){return((-3.75*v)(1-v^2)*(abs(v)<=1))}
98      v<-(t-datos)/h
99      dNbiw<-dbiw((t-datos)/h)
100     dlambdaestima<-(1/(h^2))*sum(dNbiw)
101     dlambdaestima
102 }
103
104 ##############################################################################
105 ################## Estimador rocof con núcleo Triweight ######################
106 ##############################################################################
107 Ltriweight<-function(t,datos,h){
108     triw<-function(v){return(1.09375*((1-v^2)^3)*(abs(v)<=1))}
109     v<-(t-datos)/h
110     Ntriw<-triw((t-datos)/h)
111     lambdaestima<-(1/h)*sum(Ntriw)
112     lambdaestima
113 }
114
115 ##############################################################################
116 ######## Derivada del estimador rocof con núcleo Triweight ###########
117 ##############################################################################
118 dLtriweight<-function(t,datos,h){
119     dtriw<-function(v){return((-6.5625*v)*((1-v^2)^2)*(abs(v)<=1))}
120     v<-(t-datos)/h
121     dNtriw<-dtriw((t-datos)/h)
122     dlambdaestima<-(1/(h^2))*sum(dNtriw)
123     dlambdaestima
```

```
}

#############################################################################
################## Estimador rocof con núcleo Gausiano ######################
#############################################################################
Lgaussian<-function(t,datos,h){
    gau<-function(v){return(0.39894*exp(-0.5*v^2))}
    v<-(t-datos)/h
    Ngau<-gau((t-datos)/h)
    lambdaestima<-(1/h)*sum(Ngau)
    lambdaestima
}

#############################################################################
########## Derivada del estimador rocof con núcleo Gausiano #############
#############################################################################
dLgaussian<-function(t,datos,h){
    dgau<-function(v){return((-0.39894*v)*exp(-0.5*v^2))}
    v<-(t-datos)/h
    dNgau<-dgau((t-datos)/h)
    dlambdaestima<-(1/(h^2))*sum(dNgau)
    dlambdaestima
}

#############################################################################
################## Estimador rocof con núcleo Coseno ########################
#############################################################################
Lcoseno<-function(t,datos,h){
    cosi<-function(v){return(0.25*pi*cos(0.5*pi*v)*(abs(v)<=1))}
    v<-(t-datos)/h
    Ncosi<-cosi((t-datos)/h)
    lambdaestima<-(1/h)*sum(Ncosi)
    lambdaestima
}

#############################################################################
########## Derivada del estimador rocof con núcleo Coseno #############
#############################################################################
dLcoseno<-function(t,datos,h){
    dcosi<-function(v){return((-1)*0.125*(pi*pi)*sin(0.5*pi*v)*(abs(v)
    <=1))}
    v<-(t-datos)/h
    dNcosi<-dcosi((t-datos)/h)
    dlambdaestima<-(1/(h^2))*sum(dNcosi)
    dlambdaestima
}

```

```
170 ##########################################################################################
171 ################## Estimador rocof con núcleo Logístico ##################################
172 ##########################################################################################
173 Llogistico<-function(t,datos,h){
174     logi<-function(v){return(0.5/(1+cosh(v)))}
175     v<-(t-datos)/h
176     Nlogi<-logi((t-datos)/h)
177     lambdaestima<-(1/h)*sum(Nlogi)
178     lambdaestima
179 }
180 
181 ##########################################################################################
182 ######### Derivada del estimador rocof con núcleo Logístico ##############################
183 ##########################################################################################
184 dLlogistico<-function(t,datos,h){
185     dlogi<-function(v){return(-0.5*sinh(v)/((1+cosh(v))^2))}
186     v<-(t-datos)/h
187     dNlogi<-dlogi((t-datos)/h)
188     dlambdaestima<-(1/(h^2))*sum(dNlogi)
189     dlambdaestima
190 }
191 
192 ##########################################################################################
193 ############ Se simulan los tiempos según el modelo elegido ##############################
194 ##########################################################################################
195 ### n: longitud de la muestra t1,t2,...,tn.
196 ### typefunct: forma de la función de riesgo teórica de la simulación:
197 ### (constante = "m1", creciente = "m2", periódica tipo I = "m3",
198 ### periódica tipo II = "m4").
199 simula.times<-function(s,n,typefunct){
200     ### Se simula una muestra ###
201     ### Se genera la muestra para los seis modelos ###
202     ss<-s;set.seed(ss)
203     u<-runif(n, min=0, max=1)
204     x<- -log(u)
205     r<-matrix(0,n,0)
206     r0<-0
207     r[1]<-r0+x[1]
208     for (i in 2:n){
209         r[i]<-r[i-1]+x[i]
210     }
211     vt<-matrix(0,n,0)
212 
213     ## si es el modelo 1 (constante)
214     if (typefunct=="m1")
215     {
216         for (i in 1:n){
```

```
            vt[i]<-r[i]
        }
    }
    ## si es el modelo 2 (creciente)
    if (typefunct=="m2")
    {
        for (i in 1:n){
            vt[i]<-(r[i])^(1/3)
        }
    }
    ## si es el modelo 3 (periÓdico tipo I)
    if (typefunct=="m3")
    {
        for (i in 1:n){
            vt[i]<-Inv.Lp(r[i])
        }
    }
    ## si es el modelo 4 (periÓdico tipo II)
    if (typefunct=="m4")
    {
        for (i in 1:n){
            vt[i]<-Inv.Lp2(r[i])
        }
    }
    return(vt)
}

#############################################################################
##### Función que coge las dos matrices de los intervalos de ########
##### confianza data1=Inferior, data2=Superior y las transforma #####
##### a una matriz de signos ################################################
#############################################################################
signos1<-function(data1,data2){
    signi<-matrix(0,nrow(data1),ncol(data1))
    for (i in 1:nrow(data1)){
        for (j in 1:ncol(data1)){
            if (is.na(data1[i,j]) || is.na(data2[i,j])){
                signi[i,j]<-NA
            }
            else if ((sign(data1[i,j])==sign(data2[i,j])) & (sign(
    data1[i,j])>0)){
                signi[i,j]<-1
            }
            else if ((sign(data1[i,j])==sign(data2[i,j])) & (sign(
    data1[i,j])<0)){
                signi[i,j]<--1
            }
```

```
            else if (sign(data1[i,j])!=sign(data2[i,j])){
                signi[i,j]<-0
            }
        }
    }
    return(signi)
}

#############################################################################
### Función que detecta cuantos cambios hay en la matriz de signos ##
#############################################################################
changes<-function(dat){
    limp<-rle(sign(dat))[[-1]]
    limpT<-subset(limp,limp!=0)
    cont<-sum(diff(sign(limpT))!=0)
    return(cont)
}

#############################################################################
##### Función que evalúa el número de veces que se repite cada #######
##### valor ###########################################################
#############################################################################
verepfun<-function(data){
    verep<-list(table(data[,1]),table(data[,2]),table(data[,3]),table(
    data[,4]),table(data[,5]),table(data[,6]),table(data[,7]),table(
    data[,8]),table(data[,9]),table(data[,10]),table(data[,11]))
    return(verep)
}

#############################################################################
### Función que recoge si está o no (1 o 0) contenida la derivada ###
### teórica en cada uno de las estimaciones de cada ancho de banda ##
#############################################################################
coverage<-function(data,data1,data2){
    cont<-matrix(0,nrow(data1),ncol(data1))
    for (j in 1:ncol(data1)){
        for (i in 1:nrow(data1)){
            if (is.na(data1[,j][i]) || is.na(data2[,j][i])){
                cont[i,j]<-NA
            }
            else if (data1[,j][i]<data2[,j][i]){
                cont[i,j]<-length(which((data[i] >= data1[,j][i]) & (
    data[i] <= data2[,j][i])))
            }
        }
    }
    return(cont)
```

```
306 }
307
308 ##############################################################################
309 ############## Función que divide una matriz en listas #######################
310 ##############################################################################
311 dividematriz<-function(mat,numfil,numlist){
312     it <- ihasNext(isplitRows(mat, chunkSize=numfil))
313     rt<-list()
314     while (hasNext(it)) {
315         for (s in 1:numlist){
316             rt[[s]]<-nextElem(it)
317         }
318     }
319     return(rt)
320 }
321
322
323 ##############################################################################
324 ########## El numero de muestras en las simulaciones es R ###################
325 ##############################################################################
326 R<-1000
327 ##############################################################################
328 ################ Se define el modelo que va a simularse ######################
329 ##############################################################################
330 ### Esto es: tamaño de muestra (n), modelo (typefunct),
331 ### "m1" = constante, "m2" = creciente, "m3" = periódico tipo I,
332 ### "m4" = periódico tipo II
333 typefunct<-"m1"
334 ##############################################################################
335 ##################### Se define la longitud muestral #########################
336 ##################### n: longitud de la muestra t1,t2,...,tn. ################
337 ##############################################################################
338 n<-100
339 #n<-500
340 #n<-1000
341 ##############################################################################
342 ### para una rejilla de tiempos (vector de puntos de estimacion) ####
343 ### fija de tamaño N #########################################################
344 N<-401
345 ################ número de parámetros de suavizado ###########################
346 nh<-11
347
348 ##############################################################################
349 ################### Se generan las muestras Normales #########################
350 ##############################################################################
351 ################ Se almacenan las 1000 salidas simuladas #####################
352 ##############################################################################
```

```
353 muest<-list()
354 matriz.deriv.teor<-list()
355 ICnp.inf<-list()
356 ICnp.sup<-list()
357 matriz.codigosnp<-list()
358 ICns.inf<-list()
359 ICns.sup<-list()
360 derivadas<-list()
361 matriz.codigosns<-list()
362 t0<-Sys.time()
363 for (s in 1:R)
364 {
365     ## Función que simula una muestra
366     datos<-simula.times(s,n,typefunct)
367     tt<-seq(0.0008953162,134.481,len=N)
368     ## Derivada de la función teórica
369     dfith<-sapply(tt,function(t){return(0)})#para simular el caso
    constante
370     #dfith<-sapply(tt,function(t) {return(6*t)})#creciente
371     #dfith<-sapply(tt,function(t) {return(-25*pi*sin(0.5*pi*t))})#
    periodica tipo I
372     #dfith<-sapply(tt,function(t) {return(-75*pi*sin(1.5*pi*t))})#
    periodica tipo II
373     ## con la muestra se evalúa la funcion que hace el SiZer
374     sizer.res.norm<-SiZerIFLL(datos, ikernel="epanech", alpha=0.05,
    icolor=1,
375                               nbin=401, minx= 0.0008953162, maxx
    =134.481, bpar=0, nsh=11, shmin=10, shmax=30)
376     ## 1 si es negativo
377     ## 4 si es positivo
378     ## 2 si no hay suficientes observaciones
379     ## 3 si el intervalo contiene al cero
380     matriz.deriv.teor[[s]]<-dfith
381     ## se guardan las muestras ###
382     muest[[s]]<-datos
383     #### Caso Normal ####
384     ICnp.inf[[s]]<-sizer.res.norm$icinfp # extremo inferior del
    intervalo de confianza normal puntual
385     ICnp.sup[[s]]<-sizer.res.norm$icsupp # extemo superior del
    intervalo de confianza normal puntual
386     matriz.codigosnp[[s]]<-sizer.res.norm$mapap # matriz de colores,
    ic normal puntual
387     ICns.inf[[s]]<-sizer.res.norm$icinfs # extremo inferior del
    intervalo de confianza normal simultáneo
388     ICns.sup[[s]]<-sizer.res.norm$icsups # extemo superior del
    intervalo de confianza normal simultáneo
389     derivadas[[s]]<-sizer.res.norm$deriva # estimación de las
```

```
    derivadas
390     matriz.codigosns[[s]]<-sizer.res.norm$mapas # matriz de colores,
    ic normal simultáneo
391     print(paste('Replica: ',s))
392     t1<-Sys.time()
393     print(paste('tiempo:',t1-t0))
394     t0<-t1
395 }
396
397
398 ##############################################################################
399 ##### Se guardan las 1000 bases de datos en un fichero .Rdata #######
400 ##############################################################################
401 save.image("Res_Gauss_Modelo_tamaño.RData")
402
403 ##############################################################################
404 ##### Se carga las 1000 bases de datos para operar con ellas ########
405 ##############################################################################
406 load("Res_Gauss_M1_N100.RData")
407 ls()
408
409
410 ##############################################################################
411 ################## Se generan las muestras Bootstrap ####################
412 ##############################################################################
413
414 t0<-Sys.time()
415 for (s in init:fin){
416     ## Función que simula una muestra
417     datos<-simula.times(s,n,typefunct)
418     tt<-seq(0.0008953162,134.481,len=N)
419     ## Derivada de la función teórica
420     dfith<-sapply(tt,function(t){return(0)})#constante
421     #dfith<-sapply(tt,function(t) {return(6*t)})#creciente
422     #dfith<-sapply(tt,function(t) {return(-25*pi*sin(0.5*pi*t))})#
    periódica tipo I
423     #dfith<-sapply(tt,function(t) {return(-75*pi*sin(1.5*pi*t))})#
    periódica tipo II
424     ## con la muestra se evalúa la función que hace el SiZer, Código
    en proceso de publicación en un R package
425     sizer.res.boot<-SiZerIFLL(datos, ikernel="epanech", alpha=0.05,
    icolor=1,
426                               nbin=401, minx=0.0008953162, maxx
    =134.481, bpar=0, nsh=11, shmin=10, shmax=30, B=500)
427
428     assign(paste0("ICbp.inf", s), sizer.res.boot$icinfp)
429     assign(paste0("ICbp.sup", s), sizer.res.boot$icsupp)
```

```
    assign(paste0("matriz.codigosbp", s), sizer.res.boot$mapap)
    assign(paste0("ICbs.inf", s), sizer.res.boot$icinfs)
    assign(paste0("ICbs.sup", s), sizer.res.boot$icsups)
    assign(paste0("matriz.codigosbs", s), sizer.res.boot$mapas)


    filename1 <- paste (init,'_',typefunct,"_ICbp_inf_n",n ,".txt",
    sep ="")
    filename2 <- paste (init,'_',typefunct,"_ICbp_sup_n",n ,".txt",
    sep ="")
    filename3 <- paste (init,'_',typefunct,"_matriz_codigosbp_n",n  ,"
    .txt", sep ="")
    filename4 <- paste (init,'_',typefunct,"_ICbs_inf_n",n, ".txt",
    sep ="")
    filename5 <- paste (init,'_',typefunct,"_ICbs_sup_n",n  ,".txt",
    sep ="")
    filename6 <- paste (init,'_',typefunct,"_matriz_codigosbs_n",n  ,"
    .txt", sep ="")

    write.table( get(paste0("ICbp.inf",s)) , file = filename1 , row.
    names =FALSE ,col.names = FALSE, append=TRUE)
    write.table( get(paste0("ICbp.sup",s)) , file = filename2 , row.
    names =FALSE ,col.names = FALSE, append=TRUE)
    write.table( get(paste0("matriz.codigosbp",s)) , file = filename3
    , row.names =FALSE ,col.names = FALSE, append=TRUE)
    write.table( get(paste0("ICbs.inf",s)) , file = filename4 , row.
    names =FALSE ,col.names = FALSE, append=TRUE)
    write.table( get(paste0("ICbs.sup",s)) , file = filename5 , row.
    names =FALSE ,col.names = FALSE, append=TRUE)
    write.table( get(paste0("matriz.codigosbs",s)) , file = filename6
    , row.names =FALSE ,col.names = FALSE, append=TRUE)

    print(paste('Replica: ',s))
    t1<-Sys.time()
    print(paste('tiempo:',t1-t0))
    t0<-t1
}

ICbp.inf <- read.table("All_m1_ICbp_inf_n100.txt", header=FALSE)
ICbp.sup <- read.table("All_m1_ICbp_sup_n100.txt", header=FALSE)
ICbs.inf <- read.table("All_m1_ICbs_inf_n100.txt", header=FALSE)
ICbs.sup <- read.table("All_m1_ICbs_sup_n100.txt", header=FALSE)
matriz.codigosbp <- read.table("All_m1_matriz_codigosbp_n100.txt",
    header=FALSE)
matriz.codigosbs <- read.table("All_m1_matriz_codigosbs_n100.txt",
    header=FALSE)
ICbp.inf<-data.matrix(ICbp.inf)
```

```
ICbp.sup<-data.matrix(ICbp.sup)
ICbs.inf<-data.matrix(ICbs.inf)
ICbs.sup<-data.matrix(ICbs.sup)
matriz.codigosbp <-data.matrix(matriz.codigosbp)
matriz.codigosbs <-data.matrix(matriz.codigosbs)
ICbp.inf<-dividematriz(ICbp.inf,401,1000)
ICbp.sup<-dividematriz(ICbp.sup,401,1000)
ICbs.inf<-dividematriz(ICbs.inf,401,1000)
ICbs.sup<-dividematriz(ICbs.sup,401,1000)
matriz.codigosbp <- dividematriz(matriz.codigosbp,401,1000)
matriz.codigosbs <- dividematriz(matriz.codigosbs,401,1000)


################################################################################
## LOS SIGUIENTES PROCEDIMIENTOS SE REALIZAN A MODO ILUSTRATIVO ######
## PARA EL CASO NORMAL, PARA EL BOOTSTRAP SERÍAN IDÉNTICOS ###########
################################################################################


################################################################################
############### ANÁLISIS DE LOS CAMBIOS EN LA DERIVADA ################
################################################################################

################################################################################
########### Corrección del tamaño efectivo de la muestra ##############
################################################################################

for (i in 1:R){
    ICnp.inf[[i]][matriz.codigosnp[[i]]==2]<-NA
    ICnp.sup[[i]][matriz.codigosnp[[i]]==2]<-NA
    ICns.inf[[i]][matriz.codigosns[[i]]==2]<-NA
    ICns.sup[[i]][matriz.codigosns[[i]]==2]<-NA
}

################################################################################
###### Se analizan el número de 2 hay en la matriz de códigos #######
################################################################################
contandonp<-matrix(0,R,nh)
contandons<-matrix(0,R,nh)
for (j in 1:nh){
    for (i in 1:R){
        contandonp[i,j]<-sum(matriz.codigosnp[[i]][,j] == 2)
        contandons[i,j]<-sum(matriz.codigosns[[i]][,j] == 2)
    }
}

################################################################################
```

```
### Matriz de signos de cada intervalo de confianza normal puntual ##
#########################################################################
signostodosnp<-list()
for (i in 1:R){
    signostodosnp[[i]]<-signos1(ICnp.inf[[i]],ICnp.sup[[i]])
}

#########################################################################
###### Matriz de signos de cada intervalo de confianza normal #######
###### simultáneo ###################################################
#########################################################################
signostodosns<-list()
for (i in 1:R){
    signostodosns[[i]]<-signos1(ICns.inf[[i]],ICns.sup[[i]])
}

#########################################################################
##### Se calculan los cambios (número de picos y valles juntos) #####
##### que se producen debido a los cambios en los intervalos de #####
##### confianza #####################################################
#########################################################################
cambiosnp<-matrix(0,R,nh)
for (i in 1:R){
    for (j in 1:nh){
        cambiosnp[i,j]<-changes(signostodosnp[[i]][,j])
    }
}
cambiosns<-matrix(0,R,nh)
for (i in 1:R){
    for (j in 1:nh){
        cambiosns[i,j]<-changes(signostodosns[[i]][,j])
    }
}

#########################################################################
############## Anota las veces que se repite cada valor ###############
#########################################################################
#### Caso Normal puntual ####
verepnp<-verepfun(cambiosnp)
#### Caso Normal simultáneo ####
verepns<-verepfun(cambiosns)


#########################################################################
###### ANÁLISIS DE LA COBERTURA DE LOS INTERVALOS DE CONFIANZA ######
#########################################################################

```

```
###############################################################################
##### Para cada muestra, recoge si está o no (1 o 0) contenida ######
##### la derivada teórica en cada uno de las 401 estimaciones #######
##### de cada ancho de banda ##################################################
###############################################################################
contenida1np<-list()
for (i in 1:R){
    contenida1np[[i]]<-coverage(matriz.deriv.teor[[i]],ICnp.inf[[i]],
    ICnp.sup[[i]])
}
contenida1ns<-list()
for (i in 1:R){
    contenida1ns[[i]]<-coverage(matriz.deriv.teor[[i]],ICns.inf[[i]],
    ICns.sup[[i]])
}

###############################################################################
######## Se construye la matriz que suma la cobertura de cada #######
######## posición de cada matriz de cada muestra, lo que devolverá ##
######## una única matriz para plotear y ver las coberturas ##########
###############################################################################
cover1np<-matrix(0,R,nh)
cover1ns<-matrix(0,R,nh)
cover1np<-Reduce('+', contenidanp)
cover1ns<-Reduce('+', contenidans)

###############################################################################
### Matriz que recoge por columnas de cada muestra, de la matriz ####
### contenidanp y contenidans, el mínimo valor(o sea si hay 0 o no)##
### para cada ancho de banda y los guarde en una matriz cover2np y ##
### cover2ns ##################################################################
###############################################################################
cover2np<-matrix(0,R,nh)
cover2ns<-matrix(0,R,nh)
cover2np<-lapply(contenidanp,FUN=function(x)apply(x,MARGIN=2,FUN=min,
    na.rm = FALSE))
cover2ns<-lapply(contenidans,FUN=function(x)apply(x,MARGIN=2,FUN=min,
    na.rm = FALSE))

###############################################################################
### Vector que sume por columnas a la matriz cover2np y cover2ns y ##
### calcule el porcentaje de cobertura ########################################
###############################################################################
totalTnp<-double()
totalTns<-double()
totalTnp<-((Reduce('+', cover2np))/R)*100
totalTns<-((Reduce('+', cover2ns))/R)*100
```

```
600
601 ##########################################################################
602 ###### Gráfica de las coberturas puntuales Normal y Bootstrap ########
603 ##########################################################################
604 tt<-seq(0.0008953162,134.481,len=N)## tiempos escogidos
605 vh<-seq(log10(10),log10(30),len=11)### parámetros de suavizado
    escogidos en escala logarítmica
606 cobresnp<-cover1np/10
607 cobresns<-cover1ns/10
608 win.graph()
609 lvls <- pretty(range(cobresnp, cobresns,na.rm = TRUE),20)
610 filled.contour(tt2,vh,cobresnp,color.palette=terrain.colors,levels=
    lvls, xlab='tiempo',ylab='log10(h)',key.axes = axis(4, seq(0, 100,
     by = 5)))
611 win.graph()
612 filled.contour(tt2,vh,cobresns,color.palette=terrain.colors, levels=
    lvls, xlab='tiempo',ylab='log10(h)',key.axes = axis(4, seq(0, 100,
     by = 5)))
613
614 ##########################################################################
615 ##### Gráfica de las coberturas totales puntuales y simultáneas #####
616 ##### tanto Normales como Bootstrap ######################################
617 ##########################################################################
618 win.graph()
619 plot(vh,totalTbs,xlab="log10(h)",ylab="Porcentaje (%)",
620      ylim=c(0,100),type="l", lty = 4, lwd=2, main='',las=1)
621 lines(vh,totalTns,lty=1, lwd=2)
622 lines(vh,totalTnp,lty=3, lwd=2)
623 lines(vh,totalTbp,lty=2, lwd=2)
624 legend("topleft", legend=c("S.B.","S.N.",
625                            "P.N.","P.B."),cex=1,bt="n",
626        lty=c(4,1,3,2),lwd=c(2,2,2,2))
627
628 ##########################################################################
629 ###################### Gráfica anterior en color #########################
630 ##########################################################################
631 win.graph()
632 plot(vh,totalTbs,xlab="log10(h)",ylab="Porcentaje (%)",
633      ylim=c(0,100),type="l", lty = 4, lwd=2, col = 2, main='',las=1)
634 lines(vh,totalTns,lty=1, lwd=2, col = 3)
635 lines(vh,totalTnp,lty=3, lwd=2, col = 4)
636 lines(vh,totalTbp,lty=2, lwd=2, col = 5)
637 legend("topleft",legend=c("S.B.","S.N.",
638                           "P.N.","P.B."),cex=1, bt="n",
639        col=c("red","green","blue","cyan"), lty=c(4,1,3,2),lwd=c
    (2,2,2,2))
```