

El modelo de riesgos proporcionales de Cox y sus extensiones. Impacto en Estadística y Biomedicina

por

Guadalupe Gómez Melis y Carmen Cadarso-Suárez

RESUMEN. David Cox y Bradley Efron han sido distinguidos en la edición de 2017 con el Premio Fundación BBVA Fronteras del Conocimiento en la categoría de Ciencias Básicas. Este reconocimiento es fruto de sus relevantes contribuciones metodológicas en el ámbito de la estadística y de su enorme repercusión en numerosas disciplinas tales como medicina, biología, economía o astrofísica, entre otras. Este artículo se centrará en la figura de Sir David Cox, haciendo partícipe al lector de sus aportaciones más destacadas a la estadística teórica y de sus aplicaciones en numerosas disciplinas y, muy en particular, en estudios de supervivencia. Una breve pincelada sobre su vida y trayectoria científica abre paso a la presentación del modelo de riesgos proporcionales, uno de los principales hitos científicos de Sir David Cox, cuyo origen se remonta al año 1972. Seguidamente, se hará referencia a algunas de las extensiones metodológicas con mayor relevancia aplicada, tales como los modelos multiestado, los de riesgos competitivos, o aquellas que permiten la inclusión de datos longitudinales en dicho modelo. Finalmente, se discutirá el impacto científico que ha tenido el modelo de Cox de riesgos proporcionales en múltiples disciplinas, y en particular en biomedicina, por ser la que más se ha beneficiado de dicha metodología estadística.

INTRODUCCIÓN

David Cox y Bradley Efron, dos de los más grandes estadísticos en la actualidad, han sido galardonados en 2017 con el Premio Fundación BBVA Fronteras del Conocimiento en la categoría de Ciencias Básicas, por haber desarrollado métodos estadísticos «*pioneros y enormemente influyentes*» que han demostrado ser indispensables para obtener resultados confiables en un amplio espectro de disciplinas desde la medicina, la economía y las ciencias sociales hasta la astrofísica, genómica o física de partículas [5].

Sir David Roxbee Cox nació el 15 de julio de 1924 en Birmingham (Reino Unido), se formó como matemático en St. John's College (Cambridge) y recibió su doctorado en la Universidad de Leeds (Reino Unido) en 1949. Cox se incorporó en 1988 al Departamento de Estadística de la Universidad de Oxford, donde continuó trabajando hasta su jubilación en 1994. Bradley Efron nació el 24 de mayo de 1938 en St. Paul, Minnesota (Estados Unidos), se graduó como matemático en el Instituto Tecnológico de California y se doctoró en la Universidad de Stanford en 1964. Efron es titular de

la cátedra Max H. Stein de Humanidades y Ciencias y catedrático de Bioestadística en el departamento de Biomedical Data Science de la facultad de Medicina de la Universidad de Stanford.

Ambos científicos han sido profusamente premiados a lo largo de sus carreras. Entre sus distinciones destacamos la *Guy Medal in Gold*, otorgada a ambos por la Royal Statistical Society. Además, Bradley Efron ha recibido la Medalla Nacional de Ciencia (2005), el más alto honor científico de los Estados Unidos, por su excepcional trabajo en el campo de la estadística, especialmente por proponer el método *bootstrap* de remuestreo. Cox fue distinguido en 2010 con la *Copley Medal* por la Royal Society como reconocimiento a su trabajo científico en general y, en palabras del propio Cox, es el premio del cual se siente más orgulloso, puesto que pone de manifiesto el enorme esfuerzo que sus colegas y amigos han puesto en su candidatura. Efron también ha recibido de manos de la American Statistical Association la medalla Wilks en 1990 y el premio Gottfried E. Noether en 2006. La influencia del liderazgo de Cox en la teoría y la práctica de la ciencia estadística y en la creación de un entorno en el que la estadística y sus aplicaciones han florecido han sido reconocidos con el *Marvin Zelen Leadership Award* en 1998. Por último, sin que esta lista de honores sea ni mucho menos exhaustiva, destacamos que tanto Bradley Efron como Sir David Cox son miembros de la Royal Statistical Society y *Fellows* de la American Statistical Association, y que, mientras el profesor Cox fue presidente de la Royal Statistical Society entre 1980 y 1982, el profesor Efron lo fue en 2004 de la American Statistical Association.

Tanto Cox como Efron han sido, y siguen siendo, investigadores increíblemente productivos y son conocidos por sus excelentes y relevantes contribuciones a la estadística teórica y aplicada, en general, y a la bioestadística y estadística computacional, en particular. De sus extensas obras cabe destacar el método *bootstrap* [35], una técnica de remuestreo propuesta por Efron, y el modelo de riesgos proporcionales [16], propuesto por Cox. El método *bootstrap* permite, en muchas situaciones, reemplazar cálculos teóricos por simulaciones por ordenador cuando algunos de los supuestos no se cumplen o cuando el cálculo requiere desarrollos complicados. El modelo de riesgos proporcionales relaciona el tiempo de supervivencia con las covariantes, permitiendo, por ejemplo, asociar el tabaquismo y el cáncer de pulmón.

La principal contribución de Efron, el método *bootstrap*, ya se ha tratado recientemente en la LA GACETA. Remitimos al lector al artículo *Simulación e inferencia estadística* [9] del profesor Berrendero. Por este motivo nos centramos en este trabajo en las contribuciones de Sir David Cox y principalmente en el modelo de riesgos proporcionales.

Sin embargo, antes de acabar esta introducción nos parece importante mencionar que Bradley Efron no solo se ocupó y destacó con su investigación sobre el método *bootstrap* en 1979, sino que ha aportado contribuciones muy relevantes en otros ámbitos de la estadística. Conjuntamente con Carl Morris ha estudiado los estimadores *Empirical Bayes* [40, 41] propuestos por Herbert Robbins [64], habiéndolos aplicado a datos de expresión génica de *microarrays* [38]. También ha hecho importantes contribuciones al análisis de supervivencia: ha estudiado las propiedades del estimador de Kaplan y Meier [33], profundizado en el modelo de Cox, demostrado la eficiencia

asintótica del estimador del parámetro de regresión obtenido a partir de la maximización de la verosimilitud parcial en situaciones prácticas realistas [34] y aplicado el método *bootstrap* para calcular el error estándar de un parámetro cuando los datos están censurados por la derecha [36]. Además de su extensa producción, con más de 30 artículos publicados en el *Journal of the American Statistical Association* y más de 20 en los *Annals of Statistics*, es autor de varios libros, destacando la monografía de 1982 *The Jackknife, the Bootstrap and other resampling plans* [37] y su último libro, redactado con Trevor Hastie, *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science* [39].

Cox y Efron no son solo dos gigantes de la ciencia, también son bien conocidos por su mente abierta y su proximidad. Sus oficinas están siempre abiertas a cualquier persona que tenga un problema para discutir y compartir, ellos escucharán atentamente esperando aprender algo nuevo, tener otra idea.

1. SIR DAVID COX, APORTACIONES A LA CIENCIA

Las contribuciones de Sir David Cox, tanto teóricas como aplicadas, en numerosas áreas de la estadística y de la probabilidad, son relevantes y han contribuido no solo a fundamentar mejor los pilares de la estadística, y en particular del análisis de la supervivencia, sino que también han ayudado a comprender mejor las bases de la inferencia estadística y a interpretar con mayor acierto los resultados de los análisis estadísticos. El modelo de riesgos proporcionales, ampliamente utilizado en el análisis de datos de supervivencia, es posiblemente su legado más aplicado y popular. Sin embargo no es el único; el estímulo que constantemente ha obtenido de su colaboración con científicos de diferentes áreas ha conllevado muchas otras aportaciones claras y muy productivas, no solo para los investigadores que las han usado, sino a la ciencia en general. Además, su inteligente forma de abordar en profundidad los problemas planteados por los científicos ha dado lugar a importantes contribuciones teóricas. Sin lugar a dudas, Sir David Cox puede considerarse uno de los estadísticos más importantes e influyentes del mundo, y por esta razón en numerosas ocasiones ha sido descrito como «*un gigante en el campo de la Estadística*». En esta sección cubriremos a grandes rasgos su carrera académica y nos centraremos en algunas de sus contribuciones al margen del modelo de riesgos proporcionales, que será expuesto con detalle en las siguientes secciones.

Después de que David Cox acabase sus estudios de matemáticas en St. John's College, donde en 1989 sería nombrado *Honorary Fellow*, trabajó de 1944 a 1946 en el *Royal Aircraft Establishment* en Farnborough y de 1946 a 1950 en la *Wool Industries Research Association* en Leeds. Su trabajo en la industria textil le llevó a interesarse por analizar matemáticamente el proceso de reducción de una gruesa fibra de lana a hebras o hilos finos. Su tesis doctoral, *Theory of fibre motion* (1949, Universidad de Leeds) [13], fue dirigida por Henry Daniels y Bernard Welch. En una entrevista que *Statistics Views* [67] le hizo en 2014, Cox comenta que Henry Daniels (Universidad de Birmingham: matemático, estadístico y físico) fue sin lugar a dudas el científico que más le influyó.

La carrera académica del profesor Cox ha transcurrido esencialmente en el Reino Unido. Se inició como profesor ayudante en el laboratorio de Estadística de la Universidad de Cambridge (1950–1955) y siguió su carrera en el prestigioso Birkbeck College de la Universidad de Londres, donde fue nombrado *Professor* (catedrático) en 1961. Desde 1966 hasta 1988 trabajó en el Imperial College of Science and Technology de Londres, empezando como *chair* de Estadística y asumiendo cuatro años más tarde la dirección del Departamento de Matemáticas (1970–1974). En 1988 se trasladó a la Universidad de Oxford para dirigir el Nuffield College, al tiempo que formaba parte del Departamento de Estadística. Aunque Sir David Cox se jubiló formalmente en el año 1994, sigue hoy día trabajando incansablemente desde la Universidad de Oxford.

El profesor Cox ha plasmado el fruto de sus investigaciones, reflexiones y discusiones en más de tres centenares de artículos y más de 20 libros. Sus trabajos cubren un amplio espectro de los problemas estadísticos de los últimos 70 años; en lo que sigue mencionaremos solo algunos hitos científicos de su carrera.

La experiencia temprana de Cox en la industria textil le condujo a problemas fundamentales de la estadística que ha estudiado a lo largo de su vida. En su tesis doctoral analiza matemáticamente el *drafting*, una de las operaciones esenciales en la industria textil y que consiste en producir una hebra continua fina a partir de una masa formada por muchas fibras gruesas. La tesis describe el proceso de elaboración de los hilos de lana y cómo la variabilidad de este proceso y el orden en que se llevan a cabo las diferentes partes del mismo afectan al grosor y la calidad de la hebra resultante. El estudio del efecto que diferentes tratamientos tienen en el proceso de fabricación de la lana lo lleva al diseño de experimentos [14] y a su primer libro, *Planning of Experiments* [15], publicado en 1958. Su visión exhaustiva e inclusiva de los problemas estadísticos asociados a los procesos industriales llevaron a Cox a interesarse por los procesos estocásticos. Junto con H.D. Miller (Imperial College London) recogieron este tema en *The Theory of Stochastic Processes* [24].

Desde el principio de su carrera Cox se mostró interesado por los problemas de índole teórica. En su primer trabajo, publicado en 1948 en *Biometrika* [12], deriva, de dos formas diferentes, la ley asintótica del rango y compara de forma numérica sus resultados para muestras grandes con la ley exacta, dando ya un ejemplo de cómo abordar un problema de forma útil y aplicable. Al concepto de verosimilitud, central en la inferencia estadística, el profesor Cox le ha dedicado una buena parte de su tiempo. A su artículo *Partial likelihood* [17], en el que propone una nueva definición que generaliza la verosimilitud condicional y marginal, lo han seguido una docena de contribuciones con la verosimilitud como eje central. Cox aborda el efecto del muestreo en los estadísticos derivados de la función de verosimilitud [19, 6, 7], el caso en que la matriz de información es singular [65] y la estabilidad del máximo de verosimilitud [26]. Junto con su colaboradora Nancy Reid (Universidad de Toronto) explora las propiedades y posibles extensiones de la *profile likelihood* [44, 27, 28], de la *pseudo-likelihood* [30] y de la verosimilitud condicional [32]. Con ella, además, ha escrito el libro *The Theory of the Design of Experiments* [29].

Cox ha sido un trabajador incansable que ha gozado de la colaboración de más de 100 investigadores, y ha mantenido lazos científicos con muchos de sus estu-

diantes, como por ejemplo con David Hinkley (Universidad de California, Santa Barbara) y con Valerie Isham (University College London). Cabe destacar su productiva colaboración desde 1990 con Nanny Wermuth (Chalmers University of Technology/Universidad de Gotenburgo), con quien ha publicado más de 30 artículos. Destacamos su trabajo pionero conjunto sobre Modelos Gráficos para describir de forma concisa interrelaciones complejas entre un conjunto de variables [72].

Aunque David Cox ha publicado en más de 80 revistas, su principal producción se recoge en las editadas en el Reino Unido (*Biometrika*, revistas y actas de congresos de la Royal Statistical Society, etc). Cox también ha publicado en revistas de nuestro entorno más cercano, destacando los artículos *Asymptotic theory: Some recent developments*, publicado en 1983 en la revista *Qüestió* (ahora *SORT*) [18], *The relation between theory and application in statistics (with discussion)*, aparecido en *TEST* en 1995 [21], y *A note on the robust interpretation of regression coefficients*, trabajo conjunto con M.Y. Wong publicado en 1998 en *TEST*, [73].

Acabamos este breve y no exhaustivo paseo por la carrera académica de Cox destacando tres de sus últimas contribuciones. Estos tres trabajos, además de relevantes, son un buen ejemplo de cómo Sir David Cox sigue haciendo aportaciones científicas de una gran actualidad en diversos ámbitos.

En el artículo *On some principles of statistical inference* [31], escrito con Nancy Reid, discuten algunos de los principios de inferencia estadística con el objetivo de mostrar cómo pueden ser utilizados de forma coherente para una mejor interpretación de los resultados. Con cantidades muy grandes de datos, el uso de métodos estadísticos estándar tenderá a producir estimadores aparentemente muy precisos. En *Big data and precision* [22], Cox presenta una formulación que indica al menos algunas de las formas en que el error puede acumularse a medida que aumenta el número de datos y explora las posibles implicaciones en la estimación de los coeficientes de regresión. Por último, el profesor Cox remarca en la conferencia *165th Cutter Lecture on Preventive Medicine* [23], impartida en la Harvard T.H. Chan School of Public Health, que el pensamiento estadístico proporciona un conjunto unificador de ideas generales y métodos específicos relevantes cuando existe una variación natural apreciable. Continúa insistiendo en que, para ser más fructíferas, estas ideas deben combinarse perfectamente con la materia en estudio y de ninguna manera debe considerarse el análisis estadístico como un ritual para satisfacer convenciones, evaluadores y agencias reguladoras a agregar después de que se haya realizado el trabajo serio.

2. EL MODELO DE COX DE RIESGOS PROPORCIONALES

2.1. CONCEPTOS PRELIMINARES

Tal como dice el profesor F. Downton (Universidad de Birmingham) en la discusión del aclamado artículo de Cox *Regression models and life-tables (with discussion)* [16], «*Professor Cox has given us a paper which is characteristically both elegant and useful*», y por este motivo seguiremos la fuente original, con algunas licencias en la

notación, en buena parte de la exposición para introducir y describir el modelo de riesgos proporcionales.

Consideremos una población de individuos y para cada uno de ellos observemos, o bien el tiempo hasta el fallo (muerte, etc.), o bien el tiempo hasta la censura (incluyendo aquí la pérdida de seguimiento) desde un origen bien definido que en un ensayo clínico, por ejemplo, podría coincidir con la fecha de aleatorización a cada uno de los tratamientos bajo estudio. De los individuos censurados solo sabemos que su tiempo hasta el fallo es mayor que el tiempo hasta la censura —en este artículo solamente consideramos la censura por la derecha—. Denotamos por T la variable aleatoria que representa el tiempo hasta el fallo, que puede ser continua o discreta; en esta exposición lo restringimos al caso continuo. Sea $S(t)$ la función de supervivencia correspondiente a la probabilidad de que un individuo sobreviva más de t unidades de tiempo, es decir, $S(t) = \text{Prob}(T \geq t)$, y sea $\lambda(t)$ la función de riesgo (a veces llamada función de intensidad o tasa de incidencia):

$$\lambda(t) = \lim_{\Delta t \rightarrow 0^+} \frac{1}{\Delta t} \text{Prob}(t \leq T < t + \Delta t \mid T \geq t). \quad (1)$$

Intuitivamente $\lambda(t)\Delta t$ se puede interpretar como la probabilidad de que un individuo de edad t falle en el siguiente instante. Si $\lambda(t)$ es integrable podemos expresar la función de supervivencia como

$$S(t) = \exp \left\{ - \int_0^t \lambda(u) du \right\}.$$

En lo que sigue se asume que la única información disponible sobre el tiempo de fallo potencial de un individuo censurado es que excede su tiempo de censura. La suposición de que la censura es no informativa es de mucho calado, y solo puede ser contrastada si se dispone de información complementaria apropiada.

2.2. ESTIMADOR DE KAPLAN Y MEIER

En aras de la exhaustividad y de la construcción de posteriores estimadores introducimos el estimador de Kaplan y Meier [54] basado en la descomposición como límite de un producto de probabilidades condicionadas. Supongamos que tomamos observaciones de n individuos independientes e igualmente distribuidos con funciones de riesgo $\lambda(t)$ y supervivencia $S(t)$. Denotamos los diferentes tiempos de fallo por $t_{(1)} < t_{(2)} < \dots < t_{(k)}$ y por $d_{(j)}$ el número de tiempos de fallo iguales a $t_{(j)}$ ($d_{(j)} = 1$ si no hay empates). El conjunto de individuos a riesgo en t , $R(t)$, está formado por aquellos individuos que aún no han fallado y siguen en el estudio (no han sido censurados) justo antes del momento t . Denotaremos por $n_{(j)}$ el cardinal de $R(t_{(j)})$, es decir, el número de individuos que ni han fallado ni han sido censurados en el momento $t_{(j)}$. El estimador de Kaplan y Meier de la función de supervivencia, $\hat{S}(t)$, para todo t en el rango en el que hay datos ($t \leq t_{max}$) se define [47] como

$$\hat{S}(t) = \begin{cases} 1, & t < t_{(1)}, \\ \prod_{j:t_{(j)} \leq t} \left(1 - \frac{d_{(j)}}{n_{(j)}} \right), & t \geq t_{(1)}. \end{cases} \quad (2)$$

El estimador $\hat{S}(t)$, también llamado estimador *product-limit*, es una función escalonada con saltos en los tiempos de fallo. La magnitud del salto depende, no solo del número de sucesos observados en cada tiempo de fallo, sino también del patrón de censura antes del tiempo de fallo. El estimador de Kaplan y Meier se reduce a la función de supervivencia empírica cuando no hay censura, y en ese caso tendremos, para $t \geq t_{(1)}$,

$$\hat{S}(t) = \prod_{j:t_{(j)} \leq t} \left(\frac{n-j}{n-j+1} \right).$$

Las propiedades del estimador de Kaplan y Meier fueron estudiadas, entre otros, por Bradley Efron en los *Proceedings of the Fifth Berkeley Symposium* [33]. En ese trabajo se demuestra que $\hat{S}(t)$ es autoconsistente y coincide con el estimador que se obtiene mediante un algoritmo de redistribución a la derecha. $\hat{S}(t)$ es, además, bajo ciertas condiciones de regularidad, el estimador de máxima verosimilitud en la familia de todas las posibles distribuciones [54].

2.3. EL MODELO DE REGRESIÓN DE COX

En ensayos clínicos o estudios observacionales diseñados para evaluar el efecto de una intervención en el tiempo hasta que ocurre el evento de interés, es muy importante ajustar por el efecto de otras variables (covariantes) que pueden estar asociadas con el evento en cuestión. Estas covariantes pueden ser variables demográficas, como edad o género, medidas de laboratorio, como nivel de bilirrubina o presión arterial, o de otro tipo, tales como el tiempo expuesto a una terapia previa, el tipo de enfermedad, etc. En estos casos la información que se recoge sobre cada participante se amplía con las covariantes $\mathbf{z}' = (z_1, z_2, \dots, z_p)$ donde \mathbf{z} es un vector $p \times 1$ y en toda la exposición \mathbf{z}' indica el vector traspuesto de \mathbf{z} . En esta sección nos ocupamos del caso más sencillo en el que las covariantes z_ξ ($\xi = 1, \dots, p$) se recogen al inicio del estudio y son independientes del tiempo, y para el i -ésimo individuo indicamos $\mathbf{z}'_i = (z_{1i}, z_{2i}, \dots, z_{pi})$ [48]. El modelo de Cox establece la relación entre la distribución del tiempo de fallo y las covariantes \mathbf{z} en términos del modelo

$$\lambda(t | \mathbf{z}) = \exp\{\boldsymbol{\beta}'\mathbf{z}\} \lambda_0(t) = \exp\{\beta_1 z_1 + \beta_2 z_2 + \dots + \beta_p z_p\} \lambda_0(t), \tag{3}$$

donde las covariantes tienen un efecto multiplicativo en la función de riesgo, los parámetros de regresión $\boldsymbol{\beta}$ (vector $p \times 1$) se interpretan como los logaritmos de las razones de riesgo, y la función $\lambda_0(t)$, llamada función de riesgo basal, no se especifica paramétricamente y se corresponde con la función de riesgo para el conjunto de condiciones de referencia $\mathbf{z} = \mathbf{0}$. La función $\exp\{\boldsymbol{\beta}'\mathbf{z}\}$, que podría reemplazarse por otra función no negativa $\psi(\mathbf{z})$, se interpreta como el riesgo relativo de un individuo con perfil \mathbf{z} con respecto a otro con $\mathbf{z} = \mathbf{0}$. En muchos estudios epidemiológicos en los que se compara una población expuesta a una cierta condición, por ejemplo asbestos, con otra no expuesta, la tasa de riesgos o *hazard ratio*,¹

$$\text{HR} = \frac{\lambda(t | \mathbf{z})}{\lambda_0(t)} = \exp\{\boldsymbol{\beta}'\mathbf{z}\},$$

¹Se usa habitualmente el término inglés.

es a menudo la única medida de efecto que se considera y se interpreta como la razón entre la tasa de incidencia del grupo expuesto y la tasa de incidencia del grupo no expuesto. En general para una covariante continua \mathbf{z} y dado un valor de referencia \mathbf{z}_{ref} , la *hazard ratio* se entiende como

$$\text{HR}(z, z_{\text{ref}}) = \exp\{\boldsymbol{\beta}'(\mathbf{z} - \mathbf{z}_{\text{ref}})\}. \quad (4)$$

Recomendamos la lectura del artículo de Hernán *The hazards of hazard ratios* [51] para una exposición clara sobre la interpretación del riesgo relativo y sobre el uso, a menudo incorrecto, del mismo.

Una de las aplicaciones más importantes del modelo de riesgos proporcionales se da en un ensayo clínico en el que se comparan las supervivencias de dos poblaciones. Si las funciones de riesgo para cada población son $\lambda_0(t)$ y $\lambda_1(t)$, el modelo (3) establece $\lambda_1(t) = e^\beta \lambda_0(t)$, y las funciones de supervivencia están relacionadas mediante $S_1(t) = S_0(t)e^{-\beta}$. Si se introducen otras covariantes \mathbf{z} , el modelo de Cox permite obtener una estimación del efecto del tratamiento ajustado por la presencia de las mismas. La inclusión de una covariante con alto poder predictivo, además de la variable tratamiento, comporta ventajas adicionales [45].

2.4. FUNCIÓN DE VEROSIMILITUD PARCIAL

Una vez establecido el modelo (3), y si suponemos que el interés primordial estriba en el rol que juegan los parámetros de regresión $\boldsymbol{\beta}$, la estimación de los mismos debe abordarse permitiendo que $\lambda_0(t)$ sea arbitrario. Continuamos suponiendo que hay k tiempos de fallo, $t_{(1)} < t_{(2)} < \dots < t_{(k)}$, que no hay empates $d_{(j)} = 1$ —en este caso cada tiempo $t_{(j)}$ corresponde a un individuo—, y denotamos por $\mathbf{z}'_{(j)}$ el valor que toman las covariantes del individuo que ha fallado en $t_{(j)}$. Dado que la función basal $\lambda_0(t)$ no está especificada, para estudiar la influencia de las covariantes en el tiempo de supervivencia se necesitaba una modificación de la teoría clásica de máxima verosimilitud, y con esta finalidad Cox propuso la función de verosimilitud parcial,

$$L_P(\boldsymbol{\beta}) = \prod_{j=1}^k \frac{\exp\{\boldsymbol{\beta}'z_{(j)}\}}{\sum_{l \in R(t_{(j)})} \exp\{\boldsymbol{\beta}'z_{l_j}\}}, \quad (5)$$

como base para la inferencia sobre $\boldsymbol{\beta}$, argumentando que, intuitivamente, la probabilidad de que el tiempo de fallo $t_{(j)}$ corresponda efectivamente al individuo en el que se observa el fallo, condicionada al conjunto de riesgo $R(t_{(j)})$, es el j -ésimo término en (5). $L_P(\boldsymbol{\beta})$ se basa en aquella parte de los datos que no acarrear información sobre $\lambda_0(t)$ y concretamente, para cada uno de los tiempos en que ha habido un fallo, en la identidad de los individuos que fallan en ese tiempo, condicionada al número de fallos y a la identidad de los individuos a riesgo en ese momento.

La función de verosimilitud parcial es especialmente útil cuando es mucho más simple que la verosimilitud completa, y es un buen recurso cuando el método general de máxima verosimilitud no es adecuado debido a la presencia de muchos parámetros que no son relevantes (*nuisance*). Un punto que suscita mucha discusión en

Regression models and life-tables (with discussion) [16] es la posible analogía entre la verosimilitud parcial y la verosimilitud condicional. Entre otros, Kalbfleisch y Prentice notan que diferentes verosimilitudes condicionales son compatibles. Es en [17] donde Cox define formalmente la verosimilitud parcial y demuestra que no coincide ni con la verosimilitud marginal ni con la condicional.

Trabajando con la verosimilitud parcial como habitualmente se hace con la verosimilitud completa, consideremos la función de log verosimilitud parcial $\log L_P(\boldsymbol{\beta})$, la función *score* $U(\boldsymbol{\beta})$ y la matriz de información (observada) $I(\boldsymbol{\beta})$. Se definen los promedios, con pesos exponenciales, $A_{j\xi}(\boldsymbol{\beta})$, de la covariante \mathbf{z}_ξ en la población $R(t_{(j)})$ y la covarianza entre \mathbf{z}_ξ y \mathbf{z}_η en la población $R(t_{(j)})$, $\text{Cov}_{j\xi\eta}(\boldsymbol{\beta})$, como sigue:

$$A_{j\xi}(\boldsymbol{\beta}) = \frac{\sum_{l \in R(t_{(j)})} z_{jl} \exp\{\boldsymbol{\beta}'\mathbf{z}_l\}}{\sum_{l \in R(t_{(j)})} \exp\{\boldsymbol{\beta}'\mathbf{z}_l\}},$$

$$\text{Cov}_{j\xi\eta}(\boldsymbol{\beta}) = \sum_{l \in R(t_{(j)})} \frac{z_{l\xi} z_{l\eta} \exp\{\boldsymbol{\beta}'\mathbf{z}_l\}}{\exp\{\boldsymbol{\beta}'\mathbf{z}_l\}} - A_{j\xi}(\boldsymbol{\beta})A_{j\eta}(\boldsymbol{\beta}).$$

Con esta notación,

$$\log L_P(\boldsymbol{\beta}) = \sum_{j=1}^k \boldsymbol{\beta}'\mathbf{z}_{(j)} - \sum_{j=1}^k \log \left(\sum_{l \in R(t_{(j)})} \exp\{\boldsymbol{\beta}'\mathbf{z}_l\} \right),$$

$$U_\xi(\boldsymbol{\beta}) = \frac{\partial \log L_P(\boldsymbol{\beta})}{\partial \beta_\xi} = \sum_{j=1}^k \{z_{(j\xi)} - A_{j\xi}(\boldsymbol{\beta})\}, \quad \xi = 1, \dots, p, \quad (6)$$

$$I_{\xi\eta}(\boldsymbol{\beta}) = -\frac{\partial^2 \log L_P(\boldsymbol{\beta})}{\partial \beta_\xi \partial \beta_\eta} = \sum_{j=1}^k \text{Cov}_{j\xi\eta}(\boldsymbol{\beta}), \quad \xi, \eta = 1, \dots, p. \quad (7)$$

El valor esperado de $I(\boldsymbol{\beta})$ depende de los (potenciales) tiempos de censura de los individuos que han fallado, pero en general no se dispone de esta información y se recurre a argumentos asintóticos.

2.5. INFERENCIA SOBRE $\boldsymbol{\beta}$

La inferencia —pruebas de hipótesis e intervalos de confianza— sobre los parámetros $\boldsymbol{\beta}$ se basa en el estimador $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_1, \dots, \hat{\beta}_p)$ que se obtiene maximizando la verosimilitud parcial (5) o, equivalentemente, resolviendo las ecuaciones

$$U_\xi(\boldsymbol{\beta}) = \sum_{j=1}^k \{z_{(j\xi)} - A_{j\xi}(\boldsymbol{\beta})\} = 0, \quad \xi = 1, \dots, p, \quad (8)$$

y en las propiedades asintóticas de $\hat{\boldsymbol{\beta}}$. Cox conjeturó que $L_P(\boldsymbol{\beta})$ compartía las mismas propiedades de una función de verosimilitud completa, y Tsiatis [71] fue el primero en demostrarlo. Bajo condiciones de regularidad usuales y suponiendo que

$\text{Prob}(T \geq \tau) > 0$, donde τ indica el final del estudio, existe una secuencia de soluciones $\hat{\beta}_n$ de la ecuación (8) que converge casi seguramente a β , es decir, el estimador $\hat{\beta}$ es fuertemente consistente. Además $\sqrt{n}(\hat{\beta} - \beta)$ converge en distribución a una variable aleatoria normal de media 0 y de varianza σ^2 (se omite aquí su expresión), es decir, $\hat{\beta}$ es asintóticamente no sesgado y normal. El estimador semiparamétrico $\hat{\beta}$ no es asintóticamente eficiente en la clase de todas las distribuciones posibles puesto que éstas incluyen los modelos totalmente paramétricos en los que el máximo de verosimilitud es asintóticamente eficiente. La matriz de varianzas y covarianzas de $\hat{\beta}$ se obtiene a partir de (7) substituyendo β por $\hat{\beta}$.

Para contrastar la hipótesis nula global $H_0 : \beta = 0$ se puede usar el estadístico *score*. Se demuestra que, bajo las condiciones de regularidad usuales y bajo H_0 , el vector *score* $\mathbf{U}(\mathbf{0}) = (U_\xi(\mathbf{0}), \xi = 1, \dots, p)$, donde $U_\xi(\mathbf{0}) = \sum_{j=1}^k \{z_{(j\xi)} - A_{j\xi}(\mathbf{0})\}$, es asintóticamente normal de media $\mathbf{0}$ y con matriz de varianzas-covarianzas $I(\mathbf{0}) = (I_{\xi\eta}(\mathbf{0}), \xi, \eta = 1, \dots, p)$ donde $I_{\xi\eta}(\mathbf{0}) = \text{Cov}_{j\xi\eta}(\mathbf{0})$. Por consiguiente, el estadístico $\{\mathbf{U}(\mathbf{0})\}' \{I(\mathbf{0})\}^{-1} \{\mathbf{U}(\mathbf{0})\}$ se comporta como una χ^2 con p grados de libertad. Observemos que $U_\xi(\mathbf{0})$ e $I_{\xi\eta}(\mathbf{0})$ se corresponden respectivamente con la media de la covariante z_ξ y con la covarianza entre z_ξ y z_η en la población $R(t_{(j)})$ correspondientes a un muestreo aleatorio sin reemplazamiento. La prueba del *score* es invariante bajo diferentes parametrizaciones y solo necesita hacer la maximización bajo la hipótesis nula, por lo que es idónea para saber si un modelo que depende de unos parámetros debe aumentarse o reducirse con más parámetros sin necesidad de remaximizar.

Se pueden derivar otros test estadísticos, como la prueba de la razón de verosimilitudes o el test de Wald, que usa directamente el valor de $\hat{\beta}$ y se basa en que $\sqrt{n}(\hat{\beta} - \beta)$ es aproximadamente normal. La prueba de Wald es intuitivamente conveniente porque permite presentar las conclusiones del estudio en función directamente de los parámetros de interés. Sin embargo, debido a que no es invariante bajo reparametrizaciones, puede dar respuestas absurdas si la función de verosimilitud tiene una forma inusual, por ejemplo si es multimodal.

2.6. INFERENCIA SOBRE $\lambda_0(t)$

La función de riesgo basal $\lambda_0(t)$ no aparece en la verosimilitud parcial y por consiguiente no puede estimarse a partir de las soluciones de las ecuaciones de verosimilitud (8). La estimación de $\lambda_0(t)$, y de la correspondiente función de supervivencia $S_0(t)$, es esencial para poder expresar las conclusiones de los análisis directamente en términos de la supervivencia para un grupo particular de individuos así como para derivar procedimientos gráficos sencillos para evaluar el ajuste del modelo [25]. Siguiendo el enfoque no paramétrico y suponiendo que no hay empates, el estimador de máxima verosimilitud, derivado por Breslow, para la función de riesgo acumulado $\Lambda_0(t) = \int_0^t \lambda_0(u) du$, dado $\beta = \hat{\beta}$, es

$$\hat{\Lambda}_0(t) = \sum_{t_{(j)} \leq t} \frac{1}{\sum_{l \in R(t_{(j)})} \exp\{\hat{\beta}' z_{(l)}\}}. \quad (9)$$

Una vez obtenido $\hat{\beta}$ y calculado $\hat{\Lambda}_0(t)$, el estimador de la función de supervivencia para un individuo del grupo basal ($\mathbf{z} = \mathbf{0}$), se calcula como $\hat{S}_0(t) = \exp\{-\hat{\Lambda}_0(t)\}$. Para calcular el estimador de la función de supervivencia de un individuo con perfil \mathbf{z} se usa $\hat{S}(t | \mathbf{z}) = \hat{S}_0(t)^{\exp\{\hat{\beta}'\mathbf{z}\}}$. Se demuestra que, para t fijado, $\hat{S}(t | \mathbf{z})$ se comporta asintóticamente como una normal de media $S(t | \mathbf{z})$ y con varianza finita (que aquí se omite).

2.7. REPRESENTACIÓN MEDIANTE PROCESOS DE CONTEO

En la actualidad, el tratamiento riguroso de los métodos estadísticos para analizar datos de supervivencia y para estudiar factores asociados con las tasas de ocurrencia de los eventos, en particular usando el modelo de Cox, se puede unificar mediante procesos de conteo [45]. Esta formulación, que presentamos brevemente por completitud, proporciona un enfoque basado en la teoría de martingalas que no solo es elegante sino que permite obtener por un lado expresiones sencillas para los momentos de los distintos estimadores y por otro las distribuciones asintóticas de los diferentes estimadores y tests estadísticos.

Sean T y C variables aleatorias no negativas e independientes. Los datos de supervivencia se pueden resumir mediante un par (U, δ) donde $U = \min(T, C)$ y $\delta = \mathbf{1}(T \leq C)$. Para una muestra (U_i, δ_i) , $i = 1, \dots, n$, se definen los procesos estocásticos

$$\bar{N}(t) = \sum_{i=1}^n \mathbf{1}(U_i \leq t, \delta_i = 1) = \sum_{i=1}^n N^i(t), \tag{10}$$

$$\bar{Y}(t) = \sum_{i=1}^n \mathbf{1}(U_i \geq t) = \sum_{i=1}^n Y^i(t), \tag{11}$$

donde $\bar{N}(t)$ se corresponde con el número de fallos observados antes o hasta t e $\bar{Y}(t)$ representa el número de individuos a riesgo en el momento t . Para cada individuo i , el proceso estocástico $\{N^i(t), t \geq 0\}$ es un proceso contador, adaptado a una filtración $\{\mathcal{F}(t), t > 0\}$ con $N(0) = 0$, $N(t) < \infty$, para todo $t > 0$ c. s., y cuyas trayectorias son, con probabilidad igual a 1, continuas por la derecha, constantes a trozos y con saltos de tamaño 1. La filtración $\mathcal{F}(t)$ (o historia) recoge en el tiempo t el conocimiento acumulado sobre todo lo acontecido (fallos, censuras y covariantes) a todos los individuos hasta t .

Para cada individuo i el proceso $M^i(t) = N^i(t) - A^i(t)$ es una martingala de media cero. El proceso $A^i(t) = \int_0^t Y^i(u)\lambda(u) du$, llamado compensador de $N^i(t)$, representa la función de riesgo acumulada durante el tiempo que el individuo i ha sido observado. Si se denota $\bar{A}(s) = \sum_{i=1}^n A^i(s)$ y $\bar{M}(s) = \sum_{i=1}^n M^i(s)$, se observa que $d\bar{A}(s) = \lambda(s)\bar{Y}(s)$, correspondiendo $\lambda(s)$ a la función de riesgo introducida en (1). Además, $\lambda(t)\Delta t$ representa, con esta terminología, el cambio esperado del proceso contador $N^i(t)$ en el intervalo de tiempo $[t, t + \Delta t]$, condicionado a que $Y^i(t) = 1$.

El estimador de Kaplan-Meier para la función de supervivencia introducido en (2) se reescribe como

$$\hat{S}(t) = \prod_{s \leq t} \left(1 - \frac{\Delta \bar{N}(s)}{\bar{Y}(s)} \right)$$

y puede representarse como la siguiente martingala:

$$\frac{S(t) - \hat{S}(t)}{S(t)} = \int_0^t \frac{\hat{S}(s^-)}{S(s)} \frac{\mathbf{1}(\bar{Y}(s) > 0)}{\bar{Y}(s)} d\bar{M}(s).$$

La teoría asintótica para martingalas permite demostrar de forma bastante directa que el estimador $\hat{S}(t)$ es uniformemente consistente y que $\sqrt{n}[\hat{S}(t) - S(t)]$ converge a un proceso gaussiano de media cero, con incrementos independientes y con función de varianza $S^2(t)\sigma^2(t)$ (se omite la expresión de $\sigma^2(t)$ en esta exposición).

La función *score* del modelo de Cox definida en (6) puede asimismo reexpresarse en términos de la siguiente integral estocástica:

$$\mathbf{U}(t) = \sum_{i=1}^n \int_0^t \left[\mathbf{z}_i - \frac{\sum_{j=1}^n Y^j(s) \mathbf{z}_j e^{\beta' \mathbf{z}_j}}{\sum_{j=1}^n Y^j(s) e^{\beta' \mathbf{z}_j}} \right] dM_i(s).$$

Como el término entre corchetes está acotado y es predecible, $U(t)$ es una martingala. El teorema central del límite para martingalas también se aplica aquí y asegura que $n^{-1/2}\mathbf{U}(t)$ converge a un proceso gaussiano de media cero, con incrementos independientes.

3. EXTENSIONES DEL MODELO DE COX

Desde su publicación en 1972 [16], el modelo de riesgos proporcionales (3) se ha convertido en la técnica de regresión por excelencia en estudios de supervivencia con datos censurados. El desarrollo teórico posterior de procesos de conteo y de martingalas [2, 45] ha permitido extender el modelo de Cox original en varias direcciones. Destacan, entre otros, los modelos de supervivencia con: (i) covariantes tiempo-dependientes; (ii) observaciones correlacionadas y/o múltiples; (iii) efectos flexibles no lineales de las covariantes. El enorme interés suscitado por el modelo de Cox en la comunidad científica, ha derivado en la aparición de numerosas monografías de índole teórica y práctica sobre diversas técnicas de análisis de supervivencia (ver, por ejemplo, [52, 53, 55, 56, 57, 58, 70]).

A continuación exponemos aquellas extensiones metodológicas del modelo de Cox, que han causado un mayor impacto en su aplicación práctica en diversas disciplinas, sobre todo en biomedicina.

3.1. MODELO DE COX CON COVARIANTES DEPENDIENTES DEL TIEMPO

En el transcurso de un estudio de supervivencia suelen registrarse (o monitorizarse) algunas variables del individuo a lo largo del tiempo, lo que nos conduce al

concepto de covariantes tiempo-dependientes. Básicamente, existen dos tipos de covariantes tiempo-dependientes: *exógenas* y *endógenas* [62]. Las covariantes exógenas son aquellas variables «externas» al individuo sujeto a estudio, como, por ejemplo, ciertos factores medioambientales. Las covariantes endógenas, por el contrario, son aquellas cuyos niveles dependen del propio individuo. En biomedicina, ejemplos típicos de variables endógenas son los biomarcadores de enfermedades y una gran variedad de parámetros clínicos, como pueden ser los niveles de glucosa en sangre, las transaminasas o la creatinina, entre otros.

El modelo de Cox de riesgos proporcionales (3) puede extenderse directamente al caso de covariantes exógenas. En 1984, Cox y Oakes [25] ya propusieron la parametrización del riesgo basal $\lambda_0(t) = \exp\{\boldsymbol{\alpha}'\mathbf{y}(t)\}$, donde \mathbf{y} denota al vector de covariantes dependientes del tiempo. De esta forma, el modelo de Cox se puede expresar como sigue:

$$\lambda(t | \mathbf{z}) = \exp\{\boldsymbol{\beta}'\mathbf{z}\}\lambda_0(t) = \exp\{\boldsymbol{\beta}'\mathbf{z} + \boldsymbol{\alpha}'\mathbf{y}(t)\}, \tag{12}$$

donde \mathbf{z} denota el vector de covariantes basales (no dependientes del tiempo). Este modelo (12) no permite la inclusión de covariantes endógenas, ya que éstas pueden depender, a su vez, del propio proceso de supervivencia. Esto ha llevado, en los últimos años, a un nuevo desarrollo metodológico de análisis conjunto (*joint modelling*) de supervivencia y covariantes longitudinales. Esta modelización conjunta se tratará en la sección 3.4.

3.2. CÓMO AÑADIR FLEXIBILIDAD AL MODELO DE COX

En el modelo original de Cox (3) se asume que el efecto de las covariantes continuas en la supervivencia es lineal. En caso de no linealidad se ha demostrado que esta hipótesis conduce a una estimación sesgada de los parámetros del modelo y reduce la potencia de los tests estadísticos [3, 68]. Por otro lado, la modelización lineal de un efecto que claramente no lo es puede llevar erróneamente a la suposición de que los riesgos no son proporcionales en el modelo de Cox [70]. Para tratar este problema se han propuesto en los últimos años diversas extensiones flexibles del modelo (3), con la forma general

$$\lambda(t | \mathbf{z}) = \exp\{f(\mathbf{z})\}\lambda_0(t) = \exp\{f_1(z_1) + f_2(z_2) + \dots + f_p(z_p)\}\lambda_0(t), \tag{13}$$

donde f_i son funciones suaves desconocidas de las covariantes continuas z_i , que pueden ser representadas utilizando, por ejemplo, *splines* penalizados (*P-splines*, [42]). Este modelo aditivo de Cox ha sido estudiado, entre otros, por Hastie y Tibshirani [50] en el ámbito frecuentista, y por Fahrmeir y Kneib [43] en el ámbito bayesiano. En este modelo aditivo (13), para cada covariante z , la *hazard ratio* de (4) viene dada ahora por una función que toma la siguiente expresión:

$$HR(z, z_{\text{ref}}) = \frac{\exp\{f(z)\}}{\exp\{f(z_{\text{ref}})\}} = \exp\{f(z) - f(z_{\text{ref}})\},$$

siendo z_{ref} un valor establecido de antemano.

A modo de ilustración, en la figura 1 se muestran los resultados gráficos obtenidos en el trabajo de Flores et al. [46] utilizando un modelo de Cox geoaditivo bayesiano [43]. En dicho estudio, el interés se ha centrado en la evaluación de los efectos suaves de diferentes parámetros clínicos, como los linfocitos y la hemoglobina, en la mortalidad por linfoma no-Hodgkin en Perú, así como sus diferencias espaciales según el área considerada.

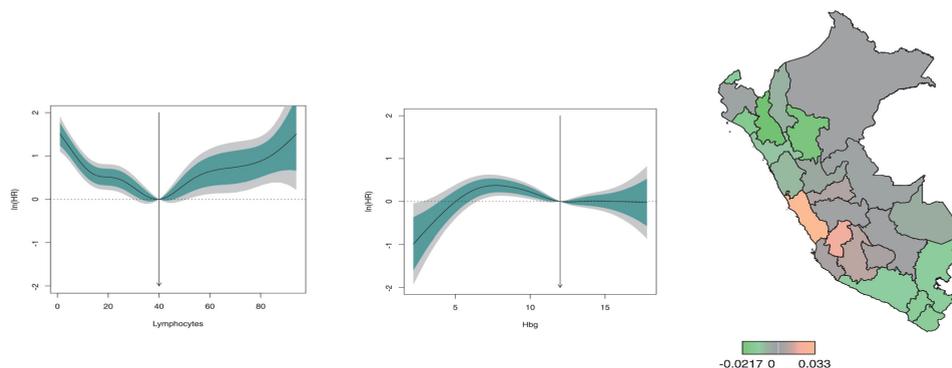


Figura 1: Estudio de mortalidad por linfoma no-Hodgkin en Perú [46]: de izquierda a derecha, función *hazard ratio* con sus correspondientes bandas de confianza al 90 y al 95 % de los linfocitos y la hemoglobina, y diferencias espaciales en la mortalidad.

3.3. MODELOS MULTIESTADO Y DE RIESGOS COMPETITIVOS

En estudios de supervivencia es habitual que los individuos experimenten varios eventos a lo largo del tiempo. En la literatura estadística se han desarrollado extensiones del modelo Cox que contemplan este tipo de situaciones, como son los modelos multiestado (*Multi-State Models*, MSM, [1, 2, 53, 59]). Los MSM son modelos para procesos estocásticos en tiempo continuo que permiten que los sujetos se puedan mover entre un número finito de estados, denominando transición al movimiento de un estado a otro. Está claro que la complejidad de un MSM depende, en gran medida, del número de estados y de las posibles transiciones entre ellos.

La estrategia más comúnmente utilizada para tratar con este tipo de situaciones prácticas, es dividir el proceso en varios modelos de supervivencia, y ajustar de forma separada el riesgo de movimiento de un estado a otro, también denominado intensidad de la transición. Cuando se cuenta con covariantes para tratar de explicar las distintas transiciones, el modelo clásicamente utilizado es el de Cox de riesgos proporcionales (3). Por norma general, se suele considerar que el proceso es markoviano o semi-markoviano [59].

El modelo multiestado más simple es el modelo de *mortalidad*, que considera únicamente dos estados: «vivo» y «muerto» y, por tanto, una única transición. Este modelo coincide con un modelo de supervivencia clásico que puede adoptar la forma

de un modelo de Cox (3). Si en lugar de considerar un único estado «vivo» se consideran k , se obtiene el denominado *modelo progresivo de k estados*. En la figura 2 se representa el esquema de un modelo progresivo de tres estados para el cáncer de mama [11], donde el estado vital de cada paciente puede ser «Viva y libre de enfermedad», «Viva con recurrencia» o «Fallecida».

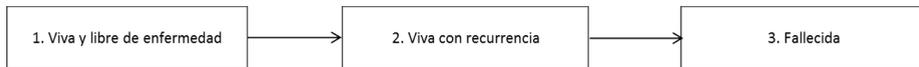


Figura 2: Esquema del modelo progresivo de tres estados en cáncer de mama [11].

Asumiendo la hipótesis de Markov, cada intensidad de transición $\lambda_{hj}(t | \mathbf{z})$ del estado h al estado j en presencia de covariantes puede ser modelada como sigue:

$$\lambda_{hj}(t | \mathbf{z}) = \exp\{\beta'_{hj}\mathbf{z}\}\lambda_{hj0}(t), \tag{14}$$

donde $h \in \{0, \dots, k - 2\}$ y $j = h + 1$, lo que se corresponde con el modelo de Cox de riesgos proporcionales (3).

En general, la implementación más usual de un modelo MSM asume riesgos proporcionales. De esta forma, todas las transiciones $h \rightarrow j$ se pueden considerar condicionalmente independientes dadas las covariantes basales y sus estados iniciales h . En consecuencia, el riesgo en cada transición se puede modelizar según (14), con $h, j \in \{0, 1, \dots, J\}$ y $h \neq j$, siendo J el número de estados.

En Beyersmann et al. [10] se presenta un claro ejemplo que corresponde a un modelo multiestado tipo *illness-death*. Se trata un estudio de supervivencia de pacientes ingresados en la Unidad de Cuidados Intensivos (UCI). En dicho estudio se analiza el impacto de la neumonía adquirida en el hospital, siendo el último estado «Salida de la UCI», tanto por fallecimiento como por alta médica. Esta situación real puede representarse a través de un modelo multiestado, cuyo esquema aparece en la figura 3.

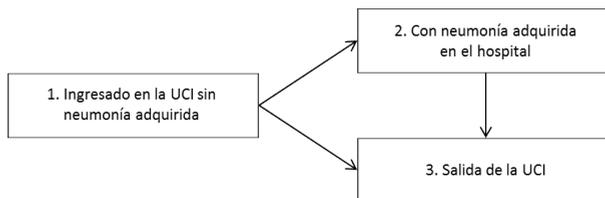


Figura 3: Esquema del modelo multiestado utilizado en [10] para analizar el impacto de la neumonía adquirida en pacientes ingresados en una Unidad de Cuidados Intensivos (UCI).

El modelo de riesgos competitivos (*Competing Risks*) es otro caso particular relevante de modelo multiestado. Este modelo ya fue planteado en 1984 por Cox y Oakes [25]. El modelo de riesgos competitivos constituye una extensión del modelo

de mortalidad simple (3), a los casos en que los sujetos pueden experimentar uno de varios eventos (o fallos) que compiten entre sí. A modo de ilustración, en la figura 4 se presenta el esquema de un modelo de riesgos competitivos considerado por Beyersmann et al. [10]. Dicho modelo corresponde a un estudio en el cual se pretende evaluar la influencia de la neumonía en la supervivencia de pacientes ingresados en la Unidad de Cuidados Intensivos. Los «fallos» compitiendo en este modelo son dos: «fallecimiento» o «alta».

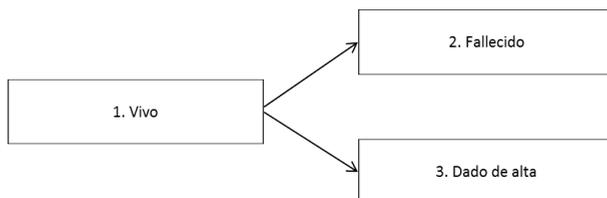


Figura 4: Esquema ilustrativo del modelo de riesgos competitivos utilizado en [10] para evaluar la influencia de la neumonía en la supervivencia de pacientes ingresados en una Unidad de Cuidados Intensivos (UCI).

En general, el riesgo debido a una causa concreta j sigue un modelo de Cox de riesgos proporcionales:

$$\lambda_{0j}(t | \mathbf{z}) = \exp\{\beta'_{0j}\mathbf{z}\}\lambda_{0j0}(t), \quad j \in \{1, \dots, J\},$$

donde J es el número de eventos o fallos distintos. En este contexto, se puede sustituir el modelo de Cox simple (3) en cada transición por alguna de sus extensiones, como por ejemplo, modelos de Cox aditivos (13).

En un modelo de riesgos competitivos, cada sujeto puede sufrir un único fallo, por lo que el concepto de dependencia es irrelevante. Por este motivo, pese a tratarse de un caso particular de MSM, se suele tratar en la literatura estadística como un modelo diferente.

3.4. DATOS LONGITUDINALES Y SUPERVIVENCIA

En los estudios biómédicos de supervivencia es habitual recoger los valores de distintos biomarcadores de un mismo paciente a lo largo del tiempo. Como ya se ha avanzado en la sección 3.1, estas medidas repetidas (conocidas como datos longitudinales) pueden no ser independientes del proceso de supervivencia. Un ejemplo claro lo podemos encontrar en el estudio presentado por Guler et al. [49], donde se ha evaluado la relación entre los niveles de ciertos parámetros, como el calcio y el fósforo, recogidos en visitas regulares cada uno o dos meses, y la muerte o el cambio de tratamiento a hemodiálisis, de pacientes sometidos a diálisis peritoneal. En la figura 5 se muestran los perfiles individuales de calcio y fósforo para todos y cada uno de los pacientes estudiados.

Para estos casos, se han propuesto modelos conjuntos de datos longitudinales y de supervivencia (*Joint Modelling for Longitudinal and Survival Data*, JMLS), que

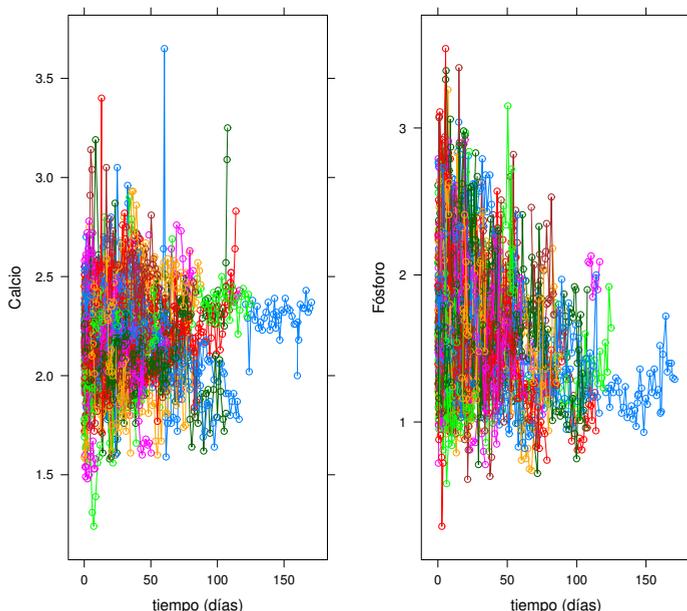


Figura 5: Perfiles individuales de calcio y fósforo de pacientes sometidos a diálisis peritoneal.

consisten básicamente, en dos submodelos: uno para los datos longitudinales y otro para la supervivencia [63], tal y como se describen a continuación:

1. *Sub-modelo longitudinal.* Generalmente, se considera el siguiente modelo lineal de efectos mixtos para la estimación del proceso longitudinal (ver [60]):

$$\begin{aligned}
 y_i(t) &= m_i(t) + \epsilon_i(t) \\
 &= \gamma' \mathbf{x}_i(t) + u_{0i} + u_{1i}(t) + \epsilon_i(t),
 \end{aligned}
 \tag{15}$$

donde $y_i(t)$ corresponde a la respuesta longitudinal del individuo i en el tiempo t , \mathbf{x}_i es el vector de diseño de los efectos fijos para el individuo i (como, por ejemplo, la edad o el sexo del sujeto) y γ son los parámetros asociados. En este modelo, u_{0i} , u_{1i} representan los efectos aleatorios para el sujeto i , que se asume siguen una distribución normal.

2. *Sub-modelo de supervivencia.* Clásicamente se considera el modelo de Cox paramétrico (3) incluyendo además las estimaciones del modelo longitudinal (15). Dicho modelo toma la siguiente expresión:

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta' \mathbf{z}_i + \alpha m_i(t)\},$$

donde $m_i(t)$ denota el verdadero valor (no observado) de la variable longitudinal en el tiempo t y α es el coeficiente asociado a $m_i(t)$.

Desde su aparición, los modelos JMLS se están extendiendo en diversas direcciones con importantes aplicaciones prácticas. Así, ya se pueden encontrar importantes contribuciones metodológicas (frecuentistas y bayesianas) en JMLS, considerando riesgos competitivos o datos longitudinales multivariados y supervivencia multivariante (ver, por ejemplo, [4, 49, 66, 69]).

4. IMPACTO DEL MODELO DE COX

Con el fin de evaluar el gran impacto que produjo —y sigue produciendo— el modelo de Cox en la literatura científica, se ha contabilizado el número total de citas del artículo original de Cox, titulado *Regression models and life-tables* [16], desde su publicación en 1972 hasta el día 15 de septiembre de 2017. En la sección 4.1 se presenta una búsqueda global en las bases de Web of Science (WoS) y Google Scholar. Dado que la medicina es el campo de mayor aplicación del análisis de supervivencia, en la sección 4.2 se realizará un estudio pormenorizado del impacto del modelo de Cox en la investigación biomédica.

4.1. RESULTADOS GLOBALES

En la base de datos WoS, el artículo de Cox [16] aparece referenciado en un total de 31 111 publicaciones, en su mayoría en forma de artículos y de *proceedings*. Utilizando Google Scholar, el número de citas es mayor, mostrando un total de 45 953 registros. En la figura 6 se representa el número acumulado de citas del artículo de Cox a lo largo de los años desde su publicación en 1972, según los registros de las dos bases de datos utilizadas. En dicha figura se puede observar que el número de citas es similar en ambas bases hasta principios de los años 90 del pasado siglo y que, a partir de ese momento, la WoS muestra un número acumulado de citas menor que Google Scholar. Esta discrepancia a partir de esas fechas puede deberse a que, a diferencia de la WoS, Google Scholar no hace distinción alguna entre el artículo original y la monografía de Cox con el mismo título [20] publicada en el año 1992.

Se ha utilizado la herramienta *Claravite Analytics* de la WoS con el objetivo de estudiar las diferentes áreas de investigación en las que se ha citado el artículo original de Cox [16]. Se ha comprobado que dicho artículo ha sido citado en un total de 132 áreas de investigación distintas, a través de 3 016 revistas científicas. En el cuadro 1 se muestran los campos de estudio en los que se engloban dichas áreas de investigación. Se puede observar, por ejemplo, un uso en 56 áreas de investigación en el campo de la biomedicina, en 21 de la biología o 15 de ingeniería.

En un intento de comparar el nivel de citación de un modelo matemático como es el de Cox, en el ámbito matemático teórico y en un campo aplicado —como es la medicina—, en el cuadro 2 se muestra el número de publicaciones de las áreas de investigación biomédicas y matemáticas con mayor número de citas a [16]. En dicho cuadro, se puede observar que oncología es el área biomédica que más utiliza el modelo de Cox, mientras que en matemáticas, tal como era de esperar, la estadística es el área donde se produce el mayor número de contribuciones metodológicas al análisis de supervivencia relacionada con el modelo Cox.

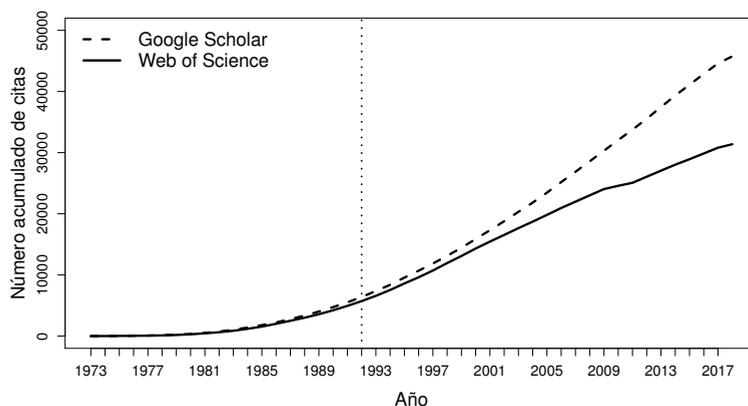


Figura 6: Número acumulado de citas del artículo de Cox [16] desde su publicación en 1972 hasta el 15 de septiembre de 2017 según los registros de la WoS (un total de 31 111 citas) y de Google Scholar (un total de 45 953 citas). La línea de referencia representa el año 1992, año de publicación de [20].

Campos de investigación	N.º de áreas de investigación
Biomedicina	56
Biología	21
Ingeniería	15
Ciencias Sociales	9
Física	7
Humanidades	5
Matemáticas	3
Derecho	3
Economía	3
Geografía	3
Geología	3
Informática	2
Meteorología	1
Química	1

Cuadro 1: Número de áreas de investigación (sobre un total de 132) por cada campo de estudio en las que se ha citado el modelo de Cox (1972), desde su aparición hasta el 15 de septiembre de 2017.

Es preciso aclarar que las áreas de investigación no son mutuamente excluyentes, en el sentido de que un artículo puede estar indexado en el *ISI Journal Citation Reports* en varias categorías a la vez. Así, un artículo publicado, por ejemplo, en la revista *Statistics in Medicine*, indexada en la categoría *Statistics & Probability*, también lo está en otras categorías, como *Medical Informatics*, *Mathematical & Computational Biology*, *Medicine Research & Experimental* o *Public Environmental*

Campo	Áreas de investigación	Número de artículos
Biomedicina	Oncología	9 305
	Cirugía	2 576
	Hematología	2 234
	Medicina General	2 188
	Salud Pública	2 162
	Cardiología	2 140
	Biología Computacional	1 417
	Radiología	1 192
	Inmunología	1 017
	Neurología	909
Matemáticas	Estadística	3 129
	Investigación Operativa	263
	Otras áreas	160

Cuadro 2: Número de artículos (sobre un total de 31 111) en las principales áreas de investigación de la biomedicina y matemáticas con mayor número de referencias al artículo de Cox (1972). Periodo: desde 1972 hasta el 15 de septiembre de 2017.

É Occupational Health.

4.2. RESULTADOS EL ÁREA DE BIOMEDICINA

Esta sección se centrará en el impacto de la metodología de Cox en la investigación biomédica. Recopilando todas las áreas de especialización médica, cabe destacar que un total de 85 718 investigadores en dichas áreas han citado al menos una vez, a lo largo de su carrera profesional, el artículo de Cox [16], publicando un total de 25 655 artículos, lo que demuestra el enorme uso de la metodología de supervivencia de Cox, desde su propuesta inicial en 1972.

A pesar del gran número de estudios biomédicos encontrados, los datos anteriores no reflejan el impacto verdadero del modelo de Cox en la investigación biomédica, infraestimando notablemente el número de citas real. Ello es debido a que una gran mayoría de publicaciones biomédicas utilizando la metodología de Cox no la referencian directamente del artículo original de Cox, sino a través de otras fuentes indirectas tales como monografías de supervivencia, manuales de software estadístico, etc.

Para realizar una mejor aproximación al impacto real del modelo de Cox en biomedicina, se ha realizado una búsqueda más pormenorizada en la WoS, y también en la base de datos PubMed (motor de búsqueda frecuente de artículos en el área biomédica). Esta búsqueda se ha llevado a cabo, ahora, con los términos que habitualmente utilizan los investigadores biomédicos para referirse al modelo de Cox en sus publicaciones: «Cox regression» OR «Cox model» OR «Proportional hazard model» OR «Proportional hazards model» OR «Proportional hazard regression» OR «Proportional hazards regression», a los que nos referiremos en el texto como

KW_Cox. Utilizando cualquiera de estos términos, las citas al modelo de Cox se incrementan notablemente, figurando en PubMed un total de 63 116 artículos, y en la WoS un total de 55 230 publicaciones. Cabe aclarar que estos datos siguen subestimando el impacto del modelo de Cox, puesto que los motores de búsqueda utilizados (WoS o PubMed) detectan las palabras clave seleccionadas, siempre y cuando figuren en el título o *abstract* del artículo, pero no en el cuerpo del artículo (algo usual en la literatura biomédica).

Basándose en los resultados de la WoS, en la figura 7 se representa el porcentaje de citas del artículo original de Cox [16], comparándola con la obtenida a través del conjunto de palabras clave habituales, KW_Cox, definido anteriormente. En dicha figura, se puede observar que a partir de 1999–2000, existe una tendencia creciente muy acusada de citaciones utilizando el conjunto de términos KW_Cox. A partir de esas mismas fechas, se detecta, en cambio, un descenso en el porcentaje de citas del artículo original. Esto es debido a que, paulatinamente, el artículo fue sustituido por referencias bibliográficas sobre la metodología original de Cox y sus extensiones, más fácilmente comprensibles por investigadores biomédicos.

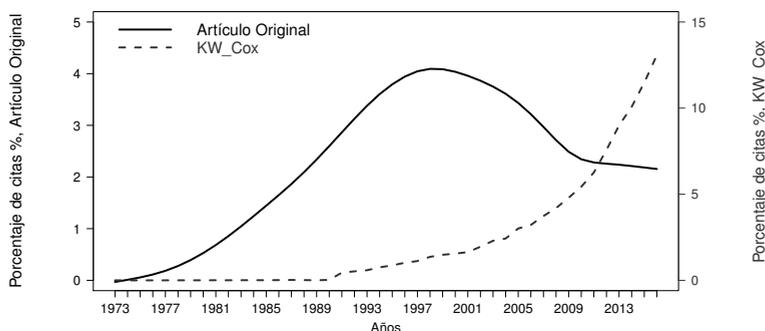


Figura 7: Porcentaje de citas por año utilizando el artículo original de Cox (total 25 304 citas) en comparación al porcentaje utilizando las palabras clave KW_Cox (total 50 387 citas) en el área de la biomedicina. Periodo: 1972–2016. Motor de búsqueda: Web of Science.

5. USO DEL MODELO DE COX: SOFTWARE ESTADÍSTICO

El uso del modelo de Cox se ha extendido enormemente a diferentes disciplinas gracias a la disponibilidad de software estadístico. Entre los programas comerciales que tienen implementado el modelo, se encuentran algunos tan conocidos como SAS (SAS Institute, Cary, NC), STATA (StataCorp, 2017), SPSS (IBM Corp., 2016), S-PLUS (TIBCO Software Inc., Palo Alto, CA) o Statistix (Analytical Software, Tallahassee, FL). También existen rutinas en MATLAB (Mathworks, Natick, MA) que permiten la estimación del modelo de Cox.

Actualmente existen programas libres, como Epi-Info (CDC, Atlanta, GA) o BayesX [8], que posibilitan la estimación de este tipo de modelos de regresión. Pero, sin duda, el entorno R (R Core Team, 2017 [61]) es el líder entre los programas informáticos que permiten el ajuste de modelos de Cox y todas sus extensiones, incluyendo las más actuales (como los MSM, el modelo de riesgos competitivos o los JLMS). En la actualidad se pueden encontrar más de 100 paquetes en CRAN [61] con los que se puede realizar este tipo de análisis de supervivencia. Entre los últimos desarrollos, se encuentran paquetes informáticos que permiten incorporar en modelos de Cox, información espacial o incluso datos funcionales entre otras covariantes predictoras.

AGRADECIMIENTOS. Guadalupe Gómez Melis y Carmen Cadarso-Suárez forman parte de la red de excelencia BIOSTATNET, proyecto MTM2015-69068-REDT (MINECO/FEDER). Las investigaciones de G. Gómez y C. Cadarso están parcialmente financiadas por los proyectos MTM2015-64465-C2-1-R y MTM2014-52975-C2-1-R, respectivamente, del Ministerio de Economía y Competitividad/FEDER, y la de G. Gómez también por el proyecto 2014 SGR 464 (GRBIO) del Departament d'Economia i Coneixement de la Generalitat de Catalunya.

REFERENCIAS

- [1] P.K. ANDERSEN, S.Z. ABILDSTROM Y S. ROSTHØJ, Competing risks as a multi-state model, *Stat. Methods Med. Res.* **11** (2002), 203–215.
- [2] P.K. ANDERSEN, O. BORGAN, R.D. GILL Y N. KEIDING, *Statistical models based on counting processes*, Springer, 1993.
- [3] G.L. ANDERSON Y T.R. FLEMING, Model misspecification in proportional hazards regression, *Biometrika* **11** (1995), 527–541.
- [4] C. ARMERO, C. FORNÉ, M. RUÉ, A. FORTE, H. PERPINÁN, G. GÓMEZ Y M. BARÉ, Bayesian joint ordinal and survival modeling for breast cancer risk assessment, *Stat. Med.* **35** (2016), 5267–5282.
- [5] BBVA, *Premio Fronteras para los estadísticos que revolucionaron el análisis de datos*. Disponible en <https://www.bbva.com/es/premio-fronteras-los-estadisticos-revolucionaron-analisis-datos/>. Acceso: 13 octubre 2017.
- [6] O.E. BARNDORFF-NIELSEN Y D.R. COX, The effect of sampling rules on likelihood statistics, *Internat. Statist. Rev.* **52** (1984), 309–326.
- [7] O.E. BARNDORFF-NIELSEN Y D.R. COX, Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator, *J. Roy. Statist. Soc. Ser. B.* **46** (1984) 483–495.
- [8] C. BELITZ, A. BREZGER, N. KLEIN, T. KNEIB, S. LANG Y N. UMLAUF, *BayesX - Software for Bayesian inference in structured additive regression models*, Version 3.0.2, 2015. Disponible en <http://www.bayesx.org>. Acceso: 13 octubre 2017.
- [9] J.R. BERRENDERO, Simulación e inferencia estadística, *Gac. R. Soc. Mat. Esp.* **18** (2015), 45–65.

- [10] J. BEYERSMANN, A. ALLIGNOL Y M. SCHUMACHER, *Competing risks and multistate models with R*, Springer, 2012.
- [11] C. CADARSO-SUÁREZ, L. MEIRA-MACHADO, T. KNEIB Y F. GUDE, Flexible hazard ratio curves for continuous predictors in multi-state models: an application to breast cancer data, *Stat. Model.* **10** (3) (2010), 291–314.
- [12] D.R. COX, A note on the asymptotic distribution of the range, *Biometrika* **35** (1948), 310–315.
- [13] D.R. COX, *Theory of fibre motion*, PhD Thesis (University of Leeds), Wool Industries Research Association, 1949. Disponible en <https://www1.maths.leeds.ac.uk/statistics/pgstats/theses/cox.pdf>. Acceso: 13 octubre 2017.
- [14] D.R. COX, Some systematic experimental designs, *Biometrika* **38** (1951), 312–323.
- [15] D.R. COX, *Planning of experiments*, Wiley, 1958.
- [16] D.R. COX, Regression models and life-tables (with discussion), *J. Roy. Statist. Soc. Ser. B.* **34** (1972), 187–220.
- [17] D.R. COX, Partial likelihood, *Biometrika* **62** (1975), 269–276.
- [18] D.R. COX, Asymptotic theory: Some recent developments, *Qüestüió* **7** (1983), 527–529.
- [19] D.R. COX, Effective degrees of freedom and the likelihood ratio test, *Biometrika* **71** (1984), 487–493.
- [20] D.R. COX. *Regression models and life-tables*, Springer, 1992.
- [21] D.R. COX, The relation between theory and application in statistics (with discussion), *Test* **4** (1995), 207–261.
- [22] D.R. COX, Big data and precision, *Biometrika* **102** (2015), 712–716.
- [23] D.R. COX, Statistical science: a grammar for research, *Eur. J. Epidemiol.* **32** (6) (2017), 465–471.
- [24] D.R. COX Y H.D. MILLER. *The theory of stochastic processes*, Methuen, 1965.
- [25] D.R. COX Y D. OAKES, *Analysis of survival data*, Chapman and Hall, 1984.
- [26] D.R. COX Y N. REID, On the stability of maximum-likelihood estimators of orthogonal parameters, *Canad. J. Statist.* **17** (1989), 229–233.
- [27] D.R. COX Y N. REID, A note on the difference between profile and modified profile likelihood, *Biometrika* **79** (1992), 408–411.
- [28] D.R. COX Y N. REID, A note on the calculation of adjusted profile likelihood, *J. Roy. Statist. Soc. Ser. B.* **55** (1993), 467–471.
- [29] D.R. COX Y N. REID, *The theory of the design of experiments*, Chapman and Hall/CRC Press, 2000.
- [30] D.R. COX Y N. REID, A note on pseudo-likelihood constructed from marginal densities, *Biometrika* **91** (2004), 729–737.
- [31] D.R. COX Y N. REID, On some principles of statistical inference, *Internat. Statist. Rev.* **83** (2015), 293–308.
- [32] A.M. CRUDDAS, N. REID Y D.R. COX, A time series illustration of approximate conditional likelihood, *Biometrika* **76** (1989), 231–237.

- [33] B. EFRON, The two sample problem with censored data, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Prentice-Hall, **4** (1966), 831–854.
- [34] B. EFRON, The Efficiency of Cox’s likelihood function for censored data, *J. Amer. Statist. Assoc.* **72** (1977), 557–565.
- [35] B. EFRON, Bootstrap methods: Another look at the jackknife, *Ann. Statist.* **7** (1979), 1–26.
- [36] B. EFRON, Censored data and the Bootstrap, *J. Amer. Statist. Assoc.* **76** (1982), 312–319.
- [37] B. EFRON, *The Jackknife, the Bootstrap, and other resampling plans*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1982.
- [38] B. EFRON, Robbins, empirical Bayes, and microarrays, *Ann. Statist.* **31** (2003), 366–378.
- [39] B. EFRON Y T. HASTIE, *Computer age statistical inference: Algorithms, evidence, and data science*, Institute of Mathematical Statistics Monographs, Cambridge University Press, 2016.
- [40] B. EFRON Y C. MORRIS, Limiting the risk of Bayes and empirical Bayes estimators. Part I: The Bayes case, *J. Amer. Statist. Assoc.* **66** (1971), 807–815.
- [41] B. EFRON Y C. MORRIS, Limiting the risk of Bayes and empirical Bayes estimators. Part II: The empirical Bayes case, *J. Amer. Statist. Assoc.* **67** (1972), 130–139.
- [42] P.H.C. EILERS Y B.D. MARX, Flexible smoothing with B-splines and penalties, *Statist. Sci.* **11** (2) (1996), 89–102.
- [43] L. FAHRMEIR Y T. KNEIB, *Bayesian smoothing and regression for longitudinal, spatial and event history data*, Oxford University Press, 2011.
- [44] H. FERGUSON, N. REID Y D.R. COX, Estimating equations from modified profile likelihood, *Estimating functions* (editado por V.P. Godambe), Oxford University Press, (1991), 279–293.
- [45] T.R. FLEMING Y D.P. HARRINGTON, *Counting processes and survival analysis*, John Wiley & Sons, 2005.
- [46] C. FLORES, M. RODRÍGUEZ-GIRONDO, C. CADARSO-SUÁREZ, T. KNEIB, G. GÓMEZ Y L. CASANOVA, Flexible geoadditive survival analysis of non-Hodgkin lymphoma in Peru, *SORT* **36** (2012), 221–230.
- [47] G. GÓMEZ, Técnicas estadísticas en análisis de supervivencia, *ICB digital [boletín electrónico]*, **5** (2002). Disponible en http://www.se-fc.org/images/icb_digital/icb05.pdf. Acceso: 13 octubre 2017.
- [48] G. GÓMEZ, O. JULIÀ Y K. LANGOHR, *Análisis de Supervivencia*, Ahlens, Barcelona, 2016.
- [49] I. GULER, C. FAES, C. CADARSO-SUÁREZ, L. TEIXEIRA, A. RODRIGUES Y D. MENDOÇA, Two-stage model for multivariate longitudinal and survival data with application to nephrology research, *Biom. J.* (2017) (en prensa).
- [50] T.J. HASTIE Y R.J. TIBSHIRANI, Exploring the nature of covariate effects in the proportional hazards model, *Biometrics* **46** (1990), 1005–1016.

- [51] M.A. HERNÁN, The hazards of hazard ratios, *Epidemiology* **21** (1) (2010), 13–15.
- [52] D.W. HOSMER JR., S. LEMESHOW Y S. MAY, *Applied survival analysis: Regression modeling of time to event data*, Second Edition, Wiley, 2008.
- [53] P. HOUGAARD, *Analysis of multivariate survival data*, Springer, 2000.
- [54] E.L. KAPLAN Y P. MEIER, Nonparametric estimation from incomplete observations, *J. Amer. Statist. Assoc.* **53** (1958), 457–481.
- [55] J.D. KALBFLEISCH Y R.L. PRENTICE, *The statistical analysis of failure time data*, Wiley-Interscience, 2002.
- [56] N. KEIDING Y P.K. ANDERSEN, *Survival and event history analysis*, Wiley, 2006.
- [57] D.G. KLEINBAUM Y M. KLEIN, *Survival analysis: A self-learning text*, Third Edition, Statistics for Biology and Health, Springer, 2012.
- [58] J. LI Y S. MA, *Survival analysis in medicine and genetics*, Chapman and Hall/CRC, 2013.
- [59] L. MEIRA-MACHADO, J. DE UÑA-ÁLVAREZ, C. CADARSO-SUÁREZ Y P.K. ANDERSEN, Multi-state models for the analysis of time-to-event data, *Stat. Methods Med. Res.* **18** (2) (2009), 195–222.
- [60] J. PINHEIRO Y D. BATES, *Mixed effects models in S and S-plus*, Springer, 2000.
- [61] R CORE TEAM, *R: A language and environment for statistical computing*, Vienna: R Foundation for Statistical Computing, 2017. Disponible en <https://www.R-project.org/>. Acceso: 13 octubre 2017.
- [62] D. RIZOPOULOS, *Joint models for longitudinal and time-to-event data: With applications in R*, CRC Press, 2000.
- [63] D. RIZOPOULOS Y P. GHOSH, A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event, *Stat. Med.* **130** (2011), 1366–1380.
- [64] H. ROBBINS, An empirical Bayes approach to statistics, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, New York, Prentice-Hall, **1** (1956), 157–163.
- [65] A. ROTNITZKY, D.R. COX, M. BOTTAI Y J. ROBINS, Likelihood-based inference with singular information matrix, *Bernoulli* **6** (2000), 243–284.
- [66] C. SERRAT, M. RUÉ, C. ARMERO, X. PIULACHS, H. PERPIÑÁN, A. FORTE, A. PÁEZ Y G. GÓMEZ, Frequentist and Bayesian approaches for a joint model for prostate cancer risk and longitudinal prostate-specific antigen data, *J. Appl. Stat.* **42** (6) (2015), 1223–1239.
- [67] STATISTICS VIEWS, An interview with Sir David Cox, *Statistics Views*, 24 Jan 2014. Disponible en <http://www.statisticsviews.com/details/feature/5770651/I-would-like-to-think-of-myself-as-a-scientist-who-happens-largely-to-specialise.html>. Acceso: 13 octubre 2017.
- [68] C.A. STRUTHERS Y J.D. KALBFLEISCH, Misspecified proportional hazard models, *Biometrika* **73** (1986), 363–369.

- [69] A-M. TANG Y N-S. TANG, Semiparametric Bayesian inference on skew-normal joint modeling of multivariate longitudinal and survival data, *Stat. Med.* **34** (2014), 824–843.
- [70] T.M. THERNEAU Y P.M. GRAMBSCH, *Modelling survival data: Extending the Cox model*, Springer-Verlag, 2000.
- [71] A.A. TSIATIS, A large sample study of Cox's regression model, *Ann. Statist.* **9** (1) (1981), 93–108.
- [72] N. WERMUTH Y D.R. COX, Graphical models: Overview, *International Encyclopedia of Social and Behavioural Sciences* (editado por N.J. Smelser y P.B. Baltes), **9** (2001), 6379–6386.
- [73] M.Y. WONG Y D.R. COX, A note on the robust interpretation of regression coefficients, *Test* **7** (1998), 287–294.

GUADALUPE GÓMEZ MELIS, DPT. D'ESTADÍSTICA I INVESTIGACIÓ OPERATIVA, UNIVERSITAT POLITÈCNICA DE CATALUNYA

Correo electrónico: lupe.gomez@upc.edu

Página web: <https://www.grbio.eu/en/about-us/cv/lupe-gomez>

CARMEN CADARSO-SUÁREZ, UNIDADE DE BIOESTADÍSTICA-FACULDADE DE MEDICINA, UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Correo electrónico: carmen.cadarso@usc.es

Página web: <http://eio.usc.es/pub/gridecmb/index.php/miembros-del-grupo/miembros-investigadores/58>