Cuadernos de Economía y Dirección de la Empresa. Núm. 19, 2004, 081-114

LUCIO FUENTELSAZ *
JAIME GÓMEZ *
YOLANDA POLO *

Aplicaciones del análisis de supervivencia a la investigación en economía de la empresa**

SUMARIO: 1. Introducción. 2. Duración y modelos econométricos tradicionales. 3. Conceptos y funciones relevantes en el análisis de supervivencia. 4. Modelos básicos en el análisis de supervivencia. 4.1. Modelos de regresión exponencial y Weibull. 4.2. Modelos de tiempo de fallo acelerado. 4.3. Modelos de riesgo proporcional. *5. Extensiones de los modelos básicos.* 5.1. Datos de duración con múltiples sucesos. 5.2. Heterogeneidad. *6. El análisis de supervivencia en Economía de la Empresa.* 6.1. Modelos de duración en economía de la empresa. 6.2. Aplicaciones del análisis de supervivencia. *7. Conclusión. Referencias bibliográficas*

RESUMEN: El objetivo de este trabajo es mostrar la aplicabilidad del análisis de supervivencia en la investigación en economía de la empresa. A pesar de que la utilización de esta técnica es relativamente común en áreas como biomedicina, ingeniería o, incluso, economía, su uso en otras disciplinas como el marketing o la organización de empresas es más bien escasa. Por tanto, el propósito del trabajo es describir las situaciones en las que el análisis de supervivencia es apropiado para la realización de análisis empíricos, presentar los modelos disponibles para el investigador y proponer ámbitos en los que su utilización es posible.

* Universidad de Zaragoza. Departamento de Economía y Dirección de Empresas. Gran Vía 2. 50005 Zaragoza (SPAIN). Teléfono: 34 976 761 000 / Fax: 34 976 761 767. E-mail: lfuente@posta.unizar.es, jgvillas@posta.unizar.es, ypolo@posta.unizar.es.

** Los autores agradecen la ayuda financiera recibida del Ministerio de Ciencia y Tecnología dentro del Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (proyectos SEC2002-01009 y SEC2002-03949) y de la Diputación General de Aragón, a través del reconocimiento de los autores del trabajo como miembros del grupo de investigación consolidado *Generés*, así como los comentarios y sugerencias de dos evaluadores anónimos. Este trabajo ha sido parcialmente realizado durante una estancia en el Departamento de Economía de la Universidad de Warwick, al que se agradece su hospitalidad. También se agradece la ayuda recibida de la Fundación Caja de Madrid

Palabras clave: análisis de supervivencia, modelos de duración, economía de la empresa, marketing, organización de empresas.

ABSTRACT: The purpose of this paper is to show the applicability of survival analysis to research performed in business economics. The use of these techniques is relatively common in areas such as biomedicine, engineering or, even, economics. Nevertheless, its employment in other disciplines such as marketing or management is scarce. Therefore, the objective of this article is to describe the situations in which survival analysis is appropriate, to present the models available to the researcher and to propose areas in which its use is possible.

Keywords: survival analysis, duration models, business economics, marketing, management.

1. Introducción

¿Son distintas las empresas pioneras de los entrantes tardíos? ¿Qué características facilitan la adopción de nuevas tecnologías? ¿Y su difusión? ¿Cómo podemos retener a nuestros clientes? ¿Y a nuestros empleados? Estas son cuestiones que se plantean habitualmente tanto en el ámbito del estudio de las empresas como en la práctica empresarial. La respuesta a las mismas tiene que ver no sólo con el hecho de si se produce un determinado suceso (entrada en el mercado, adopción de una tecnología, pérdida de fidelidad del cliente) sino también con el tiempo que transcurre hasta que dicho suceso tiene lugar. El interés de las empresas establecidas en un mercado estará en poder predecir no sólo si van a entrar o no nuevos competidores, sino también qué tipo de empresas lo van a hacer primero y cuáles después y en qué momento del tiempo se va a producir la entrada. Del mismo modo, el objetivo del regulador será encontrar y definir la estructura de mercado que proporciona más incentivos a la hora de que una nueva tecnología se difunda entre las empresas que operan en él, de modo que sus beneficios puedan llegar a ser totalmente aprovechados por el consumidor. Finalmente, uno de los objetivos de los proyectos de fidelización de consumidores es el estudio de las características de aquellos clientes que rompen más rápidamente la relación establecida con la empresa para poder diseñar planes de actuación encaminados a la retención de los mismos.

La investigación de todas estas cuestiones ha hecho uso frecuente de los métodos econométricos tradicionales: regresión lineal, o modelos logit o probit. Sin embargo, estos modelos no resultan apropiados para el análisis de situaciones en las que el interés se centra en el estudio del tiempo o duración hasta la ocurrencia de un determinado suceso. El objetivo de este trabajo es, por tanto, exponer los motivos por los cuales el análisis de supervivencia es superior a las técnicas tradicionales en el tratamiento de estas cuestiones, así como estudiar los modelos disponibles para el investigador e ilustrar el contexto en el que su uso resulta más apropiado. Si bien las aplicaciones del análisis de supervivencia son frecuentes en disciplinas como la medicina, la ingeniería o, incluso, la economía, su utilización es mucho menos habitual en el ámbito de la organización de empresas y el marketing. Este hecho es especialmente sorprendente si tenemos en cuenta el gran número de situaciones a las que esta técnica es aplicable, el incremento en el número y variedad de

modelos econométricos y la creciente disponibilidad de programas informáticos que estiman los mismos.

El trabajo se estructura como sigue. En el siguiente apartado se exponen las razones por las cuales los métodos de estimación tradicionales no son apropiados en el estudio de situaciones en las que la duración (o tiempo) hasta la ocurrencia de un determinado suceso es de interés. Tras ello, se definen, de forma general, los conceptos básicos asociados a las técnicas de análisis de supervivencia y se presentan los distintos tipos de modelos disponibles, haciendo especial referencia al modelo de riesgo proporcional propuesto por Cox. Con frecuencia, estos modelos básicos no son suficientes para captar la complejidad del fenómeno estudiado, por lo que se han propuesto extensiones que se analizan brevemente en la sección quinta. Finalmente, la sección sexta realiza una revisión de la literatura para mostrar algunas de las áreas en las que la utilización de las técnicas de supervivencia son adecuadas e ilustra la aplicación de éstas en dos de los contextos propuestos. El trabajo termina con las conclusiones.

2. Duración y modelos econométricos tradicionales

Supongamos que estamos interesados en estudiar el proceso de entrada de empresas en un determinado mercado geográfico o de producto. Es decir, nos planteamos cuestiones como las siguientes: ¿qué empresas tendrán una mayor probabilidad de entrar?, ¿cuáles lo harán antes?, ¿tienen características distintas los entrantes tempranos y los entrantes tardíos? Imaginemos también que disponemos de información sobre cada una de las posibles variables explicativas sugeridas por la literatura teórica así como sobre la ocurrencia o no del suceso que estamos estudiando y la periodicidad de la información disponible. La aproximación tradicional de la literatura empírica sobre entrada ha tendido a considerar el suceso analizado y a crear una variable ficticia que indica si éste ha tenido lugar o no a lo largo del período de observación (por ejemplo, 10 años). A continuación, se aplica un modelo logit o probit sobre el conjunto de datos para determinar la probabilidad de entrada en el período estudiado y las variables significativas para explicar la misma (ver, por ejemplo, Coterill y Haller, 1992 o Fuentelsaz y Gómez, 2001).

Es decir, la estrategia más común hubiera pasado por la estimación de un modelo como el siguiente:

$$E(Y) = \mu = \sum_{k=1}^{k} \beta_k x_k \qquad (1)$$

donde Y es una variable aleatoria cuya esperanza es μ, x_k es una variable exógena y β_k su parámetro asociado, expresión a partir de la que las especificaciones $\eta = \log[\mu/(1-\mu)]$ o $\eta = \phi^{-1}(\mu)$ (donde ϕ^{-1} es la inversa de la función de distribución normal), hubieran resultado en la estimación de un modelo logit o probit, respectivamente (Liao, 1994).

Aunque este modo de proceder puede resultar razonable, presenta varios problemas. Quizá el más importante es que una estimación como la anterior no permite aprovechar la información sobre el momento del tiempo en que se produce la entrada. Es decir, dos empresas con entradas en distintos momentos (por ejemplo, una al principio del período de observación y otra al final de dicho período) serían iguales desde el punto de vista del modelo. En ambos casos, la observación de la entrada llevaría a que el indicador apareciera con un valor igual a uno, con la imposibilidad de discriminar entre entrantes tempranos y entrantes tardíos. Una consecuencia inmediata es que los estimadores así obtenidos serían ineficientes, es decir, tendrían una varianza mayor que la resultante de la aplicación de un modelo de duración (Box-Steffensmeier y Jones, 1997; Chung, Schmidt y Witte, 1991)[1].

Un segundo inconveniente asociado a la aplicación de modelos logit o probit al estudio de cuestiones como las planteadas es que no permiten utilizar la información sobre la evolución de las variables explicativas en el tiempo[2]. Es decir, la aplicación de este tipo de técnicas desprecia la información contenida en la dimensión longitudinal de los datos. Siguiendo con el ejemplo planteado, una de las variables propuestas en la literatura para explicar tanto la probabilidad como el momento de entrada de una empresa en un mercado es el tamaño del entrante potencial. Se argumenta que las empresas más grandes son las primeras en entrar en nuevos mercados, debido a la posesión de más recursos de holgura (Cyert y March, 1963; Thompson, 1967; Haveman, 1993) y una mayor capacidad para superar las barreras a la entrada (Bain, 1956; Fuentelsaz, Gómez y Polo, 2002). Puesto que es previsible que el tamaño de una empresa varíe en el tiempo, sería aconsejable la utilización de una metodología que pudiera incorporar dicha variación al análisis.

Dado que nuestro objetivo no es sólo conocer el tipo de empresa que entra en un mercado en un intervalo de tiempo, sino también el momento en que se realiza dicha entrada, una aproximación alternativa a la utilización de los modelos de elección discreta pasaría por considerar como variable dependiente el tiempo transcurrido —o duración— (puede ser el momento en el que aparece un producto o la fecha de liberalización de un mercado geográfico). Así, si Y_i fuera la duración correspondiente a la empresa i (tiempo, medido en días, meses, años u otras escalas), esta segunda estrategia estimaría un modelo de regresión lineal que vendría dado por la expresión (Liao, 1994):

[1] El término «análisis de supervivencia» tiene su origen en la medicina y en los estudios que tratan de analizar el tiempo desde la aplicación de un determinado tratamiento hasta el suceso de interés, la muerte o la curación (supervivencia), de un enfermo. Sin embargo, como se ha señalado con anterioridad, estas técnicas también han sido utilizadas en otras disciplinas, en las que se las conoce con distintos nombres. Así, los economistas suelen hablar de «modelos de duración», los ingenieros de «técnicas de fiabilidad» o de «análisis del tiempo hasta el fallo» y los sociólogos de «análisis de la historia de un suceso». En este artículo todos estos términos son utilizados de forma indistinta.

[2] En el contexto del análisis de supervivencia, las variables explicativas suelen recibir la denominación de covariables.

$$\eta = \mu \tag{2}$$

Si bien esta estrategia tiene en cuenta el orden en el que se produce la entrada, sigue sin considerar la información contenida en la dimensión longitudinal de los datos[3, 4]. Por tanto, aunque cualquiera de las formas de proceder reseñadas puede resultar razonable y se utiliza con cierta frecuencia, ambas desprecian parte de la información disponible (Allison, 1984).

Por último, los métodos tradicionales poseen un inconveniente adicional: son incapaces de aprovechar la información contenida en las observaciones censuradas. Con este término se hace referencia a las observaciones cuya duración total no se recoge en el estudio. O, en otras palabras, aquellas observaciones para las que no se tiene información acerca de su fecha de «nacimiento», la fecha de su «muerte» o ninguna de las dos. La censura puede aparecer por distintos motivos, pero quizás el más frecuente se deriva de la existencia de períodos de observación limitados.

Para ilustrar este hecho, la Figura 1, adaptada de Yamaguchi (1991) muestra los casos de censura más habituales[5]. Como se puede comprobar, todas las observaciones son objeto de estudio desde t_0 hasta t_1, donde ambos momentos del tiempo (t_0 y t_1) se determinan con independencia de los individuos observados. Las líneas continuas horizontales representan el período de riesgo[6] para cada individuo y un «1» al final significa que se ha producido el suceso estudiado (por ejemplo, la entrada de una empresa en el mercado).

Con el fin de interpretar adecuadamente cada situación, seleccionamos dos observaciones tipo recogidas en la figura 1. Supongamos que el problema que analizamos a fecha t_1 es la entrada de una empresa en un determinado mercado geográfico a partir del momento de liberalización de ese mercado (momento t_0). La observación de la empresa A no está sometida a censura por lo que su seguimiento se realiza sin problemas durante todo el período de observación. El caso de la empresa B es diferente, porque aunque podemos seguir el com-

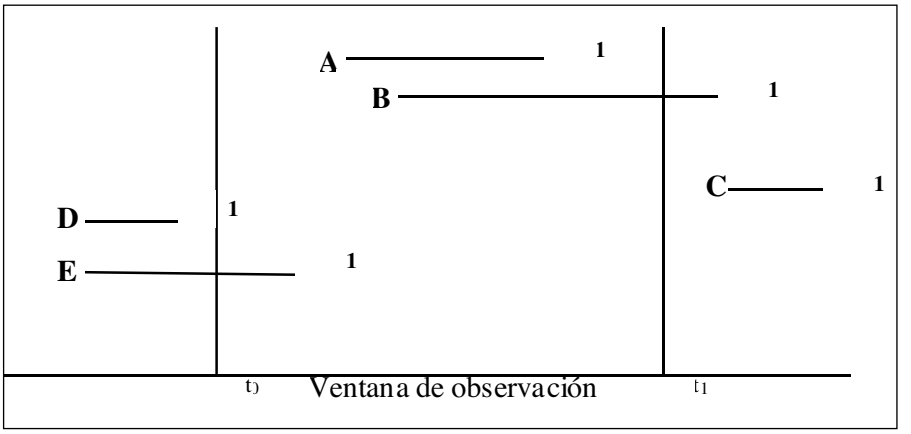
[3] Como vemos, este comentario es extensivo a modelos de regresión aplicables a variables dependientes limitadas, como el caso del modelo Tobit (dado el carácter de la variable dependiente, que toma valores no negativos, estos modelos serían más apropiados que el modelo de regresión lineal general que se discute aquí sólo a efectos ilustrativos).

[4] Podría argumentarse que la incorporación de la dimensión longitudinal sí que existe en aquellos casos en los que los datos tienen estructura de panel. Sin embargo, aunque los modelos que utilizan datos de panel suponen una mejora en términos de utilización de los datos disponibles, pueden conducir a conclusiones erróneas sobre las tasas y el momento de ocurrencia de un suceso debido a que, como señalan Singer y Spilerman (1976a, 1976b) el proceso de cambio modelizado suele ser consistente con muchas especificaciones y las conclusiones sobre las tasas y el momento en el que se produce el cambio son dependientes del intervalo temporal elegido en el diseño del panel (Box-Steffensmeier y Jones, 1997).

[5] El tratamiento de las observaciones censuradas que se realiza en esta sección está parcialmente basado en Flavián, Gómez, Polo y Martínez (1998).

[6] Aunque este concepto se define en la siguiente sección, se entiende por período de riesgo aquel en el que un determinado individuo puede experimentar el suceso de interés (es decir, aquel período en el que el individuo está «en riesgo» de que le ocurra el suceso).

FIGURA 1.— *Observaciones censuradas a derecha e izquierda*



portamiento de la misma desde su constitución, no hemos podido medir el momento de entrada ya que nuestro período de observación finaliza antes de producirse esta entrada. Lo único que podemos afirmar es que esta empresa no ha entrado en el mercado antes de t_1, pero no podemos anticipar nada acerca de su comportamiento en el futuro, por lo que diremos que la observación está *censurada a la derecha*. La interpretación del resto de casos sería similar. Así, los individuos C y D presentan períodos de riesgo fuera de la ventana de observación y están, por tanto, completamente censurados (a la derecha y a la izquierda, respectivamente). El caso del individuo E es simétrico al de B, pero en este caso la observación está *censurada a la izquierda*.

Desde nuestro punto de vista, el caso más interesante es aquél que presenta censura a la derecha (B en la figura). Este tipo de censura es frecuente en los estudios de ciencias sociales, y más concretamente, en economía y dirección de empresas. En las muestras empleadas para la realización de dichos estudios, suele ser común que alguno de los individuos no haya experimentado el suceso de interés (puede ser que al final del período de observación algunas de las empresas no hayan entrado en el mercado, la tecnología no haya sido totalmente adoptada o el cliente no haya roto su relación). En este contexto, las técnicas tradicionales tienden, de nuevo, a despreciar la información contenida en las observaciones censuradas, dado que no se considera la probabilidad de fallo después del cierre de la ventana de observación, lo cual puede sesgar los resultados obtenidos (Allison, 1984)[7]. Este problema se minimiza al utilizar el análisis de supervivencia.

[7] Conviene destacar que el tratamiento que las técnicas tradicionales ofrece a este tipo de problemas es inapropiado. En el caso del análisis logit (probit), una situación como la correspondiente al individuo B daría lugar a una observación para la cual el suceso no habría tenido lugar (por tanto, en este caso, la variable ficticia que recoge la ocurrencia del suceso aparecería con un valor igual a «cero»). En el caso del análisis de regresión tradicional, la inexistencia de un indicador de censura impide distinguir entre las observaciones censuradas y aquellas para las que el suceso ha tenido lugar.

3. Conceptos y funciones relevantes en el análisis de supervivencia

Por análisis de supervivencia entendemos un conjunto de conceptos, herramientas y técnicas dirigidas al estudio del tiempo que transcurre hasta la ocurrencia de un suceso. Como ya se ha señalado, los modelos de duración presentan ciertas ventajas sobre los modelos tradicionales de regresión ya que permiten responder a cuestiones en las que el tiempo es la variable dependiente, incluyendo en el análisis covariables dependientes del tiempo y tratando, de forma correcta, las observaciones censuradas.

Como ya se ha podido intuir, el concepto de suceso (evento) es fundamental en el análisis de supervivencia. Un evento se define como un conjunto de cambios cualitativos que tienen lugar en un determinado momento del tiempo. De acuerdo con Allison (1984), estos cambios pueden ser considerados sucesos en la medida en la que supongan una variación brusca con respecto a la situación que precede al cambio. En consecuencia, un evento puede ser, por ejemplo, la adopción de una nueva tecnología, la salida (o entrada) de una empresa de un mercado, la marcha de un empleado o el fin de la relación comercial establecida con un cliente.

Al evaluar un suceso debemos distinguir entre el período de tiempo durante el cual el evento no tiene lugar y aquél en el que ya ha sucedido. Dentro del primer período suele diferenciarse entre *período de riesgo* (durante este tiempo el evento podría producirse aunque finalmente no es así) y el *período sin riesgo* (es imposible que el evento tenga lugar). En el contexto de las decisiones de entrada, por ejemplo, el período de riesgo comienza en el momento en que la empresa analizada tiene la posibilidad de operar en el nuevo mercado. Si el producto todavía no se ha creado o el mercado está regulado y no se permite la entrada, la empresa no está todavía en el período de riesgo. Si se estudia la adopción de una nueva tecnología en una industria, el período de riesgo se inicia cuando la innovación está disponible para su empleo. En estos casos, se dice que las empresas que pueden experimentar el suceso analizado pertenecen al *conjunto de riesgo,* que se define como el grupo de individuos que pueden experimentar el evento en un momento dado. Como es obvio, el conjunto de riesgo varía a lo largo del horizonte temporal considerado.

Por otra parte, en el análisis de supervivencia hay tres distribuciones que resultan especialmente relevantes: la función de supervivencia, la función de densidad de probabilidad y la función de riesgo[8]. Sea $T>0$ una variable aleatoria que mide el tiempo hasta la ocurrencia de un determinado suceso. La *función de supervivencia* especifica la probabilidad acumulada de que el suceso objeto de estudio tenga lugar después de t. En otras palabras, la probabili-

[8] Es importante reseñar que a lo largo del artículo sólo se hace referencia a distribuciones y modelos de supervivencia continuos. Sin embargo, la literatura también ha desarrollado métodos para el análisis de datos de duración en los que bien la ocurrencia del suceso sólo tiene lugar en momentos discretos del tiempo (modelos para datos discretos) o sólo es observada en dichos momentos.

dad de que el individuo «sobreviva» hasta t, dado que no ha «muerto» hasta entonces. En el caso en el que la información disponible sea continua, la función de supervivencia, $S(t)$, puede especificarse del modo siguiente:

$$S(t) = P\ (T \geq t) \tag{3}$$

donde $S(t)$ es una función continua monótona no creciente, con S(0)=1 y $\lim_{t\to\infty} S(t)=0$ (Kalbfleisch y Prentice, 1980).

A partir de esta expresión es posible definir otras dos funciones cuyo uso quizás resulte más familiar. La función de distribución, $F(t)$, representa la probabilidad acumulada hasta t de que ocurra el suceso analizado. Matemáticamente,

$$F(t) = 1 - S(t) \tag{4}$$

Evidentemente, la función de distribución y la de supervivencia son formas equivalentes de especificar la distribución de una variable (Kiefer, 1988). Alternativamente, podemos expresar la función de distribución como:

$$F(t) = \int_0^t f(u)du \tag{5}$$

donde $f(t)$ es nuestra segunda función de interés, la *función de densidad* de la duración, cuya expresión y relación con las dos funciones anteriores es la siguiente

$$f(t) = \lim_{\Delta t\to 0^+} \frac{P\ (t \leq T < t + \Delta t)}{\Delta t} = \frac{-ds\ (t)}{dt} = \frac{dF\ (t)}{dt} \tag{6}$$

Como se puede observar, la función de densidad nos determina la probabilidad instantánea y no condicionada de que el suceso de interés tenga lugar en un momento t. Si bien tanto esta función como la de supervivencia podrían utilizarse para modelizar el tiempo hasta la ocurrencia de un suceso, la función más frecuentemente empleada es la *función de riesgo*, $h(t)$, cuya expresión es la siguiente:

$$h(t) = \lim_{\Delta t\to 0^+} \frac{P\ (t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \frac{f\ (t)}{S(t)} \tag{7}$$

La función de riesgo especifica la probabilidad condicional instantánea de que un individuo experimente el suceso («muera») en el intervalo $[t + \Delta t]$, dado que no lo ha experimentado («ha sobrevivido») hasta el momento t. En el contexto de las decisiones de entrada, representaría la probabilidad de que una empresa diversifique hacia un mercado determinado en t, dado que no lo ha hecho hasta entonces. En una interpretación alternativa la función expresa el riesgo de que la duración termine en ese mismo intervalo. Por ejemplo, en

el estudio de la duración de la relación de un cliente con la empresa, la función de riesgo expresa la probabilidad de que dicha relación tenga una duración t.

Las razones por las que se opta mayoritariamente por la función de riesgo al estudiar el tiempo hasta que ocurre un suceso son varias (Box-Steffensmeier y Jones, 1997). Quizá la más importante ya se ha apuntado, y es que dicha función tiene una interpretación inmediata en términos del riesgo de ocurrencia de un determinado evento o de finalización de una determinada relación o actividad, con utilidad en distintas aplicaciones[9].

Una segunda ventaja surge de la consideración de las propiedades de la función de riesgo (Box-Steffensmeier y Jones, 1997), ya que sólo necesitamos conocer ésta para calcular tanto la función de supervivencia como la función de densidad. A partir del concepto de función de riesgo podemos definir la función de riesgo acumulado, $H(t)$,

$$H(t) = \int_0^t h(u)du \tag{8}$$

y, utilizando la ecuación (7), podemos definir tanto la función de supervivencia como la función de densidad a partir, únicamente, de la función de riesgo[10]:

$$S(t) = \exp(-H(t)) \tag{9}$$

$$f(t) = h(t)\exp(-H(t)) \tag{10}$$

La figura 2 representa gráficamente las cinco funciones definidas para el caso de la distribución exponencial. Como podemos observar, tanto la función de riesgo acumulado como la función de distribución son monótonas crecientes, mientras que la función de supervivencia es monótona decreciente. Sin embargo, las funciones de riesgo y densidad pueden presentar distintas formas respecto a la duración. Generalmente, dicha dependencia se suele definir sobre la función de riesgo. Así, se dice que existe dependencia positiva de la duración cuando $\partial h(t)/\partial t > 0$, mientras que dicha dependencia es negativa

[9] Además de las ya mencionadas (decisiones de entrada en nuevos mercados, Fuentelsaz *et al.*, 2002, adopción y difusión de innovaciones, Hannan y McDowell, 1984, duración de la relación empresa-cliente Shaoming, 1995), otros contextos en los que se ha analizado el riesgo de ocurrencia de un suceso son el estudio de la rivalidad entre empresas, en términos de acciones y reacciones (Chen, 1996; Young *et al.*, 2000), la rotación de empleados (Hoverstad *et al.*, 1990; Morita, Lee y Mowday, 1993), la supervivencia de empresas (Luoma y Laitinen, 1991; Suarez y Utterback, 1995), la duración de las alianzas o la supervivencia de nuevos productos (Asplund y Sandin, 1999).

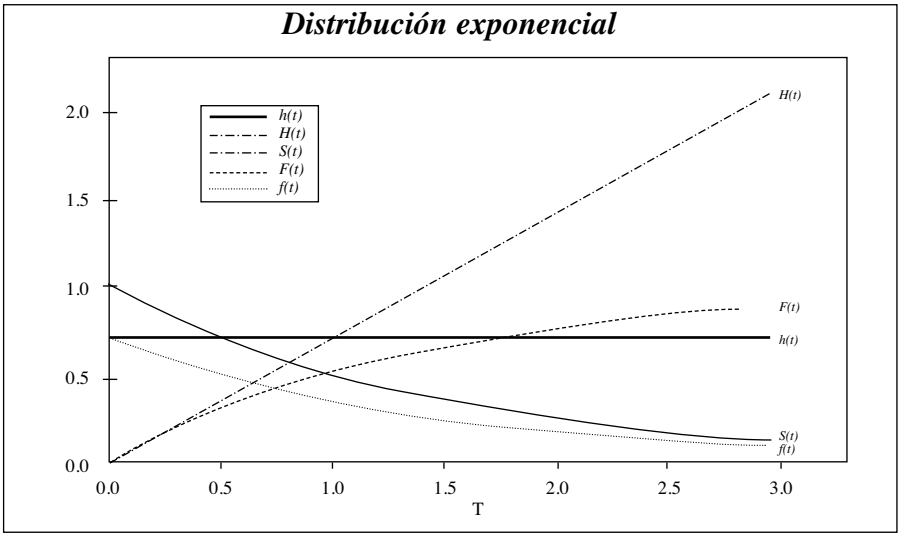
[10] Una forma de ver como se obtiene la relación (9) es la siguiente. Si tomamos logaritmos neperianos sobre la función de supervivencia y consideramos la derivada parcial de ésta respecto del tiempo, tendremos lo siguiente:

$$\frac{\partial \ln[S(t)]}{\partial t} = \frac{1}{S(t)} \frac{\partial S(t)}{\partial t} = \frac{1}{S(t)} \frac{\partial}{\partial t}[1 - F(t)] = \frac{-f(t)}{S(t)} = -h(t)$$

de donde podemos obtener que $H(t) = -\ln[S(t)]$ y $S(t) = \exp(-H(t))$.

cuando $\partial h(t)/\partial t < 0$ (Kiefer, 1988). En el caso al que hace referencia la figura 2, el riesgo de que el suceso de interés tenga lugar (el riesgo de que una duración sea igual a *t*) es una constante que no muestra dependencia del tiempo que haya durado el intervalo medido. Sin embargo, en algunas aplicaciones puede ser más correcto asumir que la función de riesgo muestre una determinada dependencia funcional respecto de la duración. Por ejemplo, la literatura sobre difusión de innovaciones sugiere la existencia de efectos epidémicos en la adopción de nuevas tecnologías por parte de las empresas establecidas en una industria. Si esto es cierto, deberíamos asumir que conforme se incrementa el tiempo desde la aparición de la innovación, el riesgo de adopción debería incrementarse y, por tanto, la función de riesgo propuesta debería mostrar dependencia positiva (Karshenas y Stoneman, 1993). Por el contrario, en otras aplicaciones la dependencia puede resultar ser negativa. Así, la evidencia empírica sobre supervivencia de nuevas empresas sugiere que el riesgo de desaparición de una empresa se reduce conforme su edad se incrementa[11]. Como veremos en la sección siguiente, los modelos de duración que se han propuesto en la literatura son lo suficientemente variados como para adaptarse a la mayoría de las hipótesis sobre la forma de la función de riesgo.

FIGURA 2.— *Funciones de distribución en el análisis de supervivencia*



[11] Algunos autores apuntan que la dependencia de la función de riesgo respecto de la duración puede ser interpretada como una señal de mala especificación del modelo. Dicha dependencia podría evitarse incluyendo en el modelo las variables que la provocan (Box-Steffensmeier y Jones, 1997; Beck 1998).

4. Modelos básicos en el análisis de supervivencia

Los métodos tradicionalmente utilizados en análisis de supervivencia pueden dividirse en tres grupos: *paramétricos, semiparamétricos* y *no paramétricos*. La principal diferencia entre ellos radica en que se tenga o no en cuenta en la estimación el efecto de las covariables sobre la función de riesgo. Los primeros (paramétricos y semiparamétricos) especifican la relación entre el riesgo de experimentar el suceso y las covariables. Los no paramétricos, sin embargo, no plantean una relación concreta entre las covariables y la variable dependiente. Puesto que el principal interés en el área de economía de la empresa suele ser la estimación del efecto de un grupo de regresores sobre la variable dependiente, nuestra discusión en este apartado se orientará hacia la descripción de los modelos que tienen en cuenta la información asociada a las observaciones incluidas en la muestra y su influencia sobre los tiempos de supervivencia.

Los procedimientos para la inclusión de covariables y la estimación de los distintos modelos propuestos para el estudio de datos de duración son análogos a los de la regresión no lineal. El punto de partida es la construcción de la función de verosimilitud[12]. En ausencia de observaciones censuradas, la función de verosimilitud de una muestra de n duraciones de longitud t_i es

$$L = \prod_{i=1}^{n} f(t_i, x) \tag{11}$$

donde $f(t_i, x)$ es la función de densidad de duración, que se supone conocida, y representa el vector de covariables, que puede contener regresores que varíen con el tiempo[13].

Como se ha señalado al comparar el análisis de supervivencia con los métodos de estimación tradicionales, una de las ventajas del primero es que consigue ofrecer un correcto tratamiento en presencia de observaciones censuradas. Si bien los métodos tradicionales tienden a desechar la información proporcionada por las observaciones cuya duración aún no ha finalizado, los

[12] Como se explicará al considerar los distintos tipos de modelos, una alternativa a la utilización de la función de máxima verosimilitud es la construcción de lo que se denomina función de verosimilitud parcial, tal y como propone Cox (1972, 1975).

[13] Como ya se ha señalado, una de las ventajas del análisis de supervivencia es que permite aprovechar la información longitudinal de las covariables cuando ésta está disponible. Aunque los modelos de duración permiten la inclusión de variables explicativas que muestran variación temporal, éstas complican la construcción de la función de verosimilitud. Dado que el análisis de estas cuestiones excede del propósito de este artículo, se ha decidido mantener la notación lo más sencilla posible, omitiendo la referencia a dicha variación temporal desde un punto de vista técnico. El lector interesado puede acudir a libros de referencia como los de Kalbfleisch y Prentice (1980, págs. 122-127), Lancaster (1990, cap. 2) o Cox y Oakes (1984). Petersen (1985, 1995) también ofrece una discusión sobre las mismas.

modelos de duración incorporan dicha información en el procedimiento de estimación a partir de la utilización de la función de supervivencia definida en la sección anterior. Es decir, en presencia de observaciones censuradas, la función de verosimilitud recoge la información de que para dichas observaciones el suceso no ha tenido lugar (o, en otras palabras, han sobrevivido) hasta un determinado momento del tiempo de la siguiente forma:

$$L = \prod_{i=1}^{n} [f(t_i,x)]^{C_i} \quad [S(t_i,x)]^{1-C_i} \tag{12}$$

donde $S(t_i,x)$ es la función de supervivencia, C_i es un indicador que toma el valor 1 si se ha observado el suceso y 0 si se ha producido censura y t_i, $f(t_i,x)$, x se han definido con anterioridad.

Como vemos, la aparición de observaciones censuradas complica la construcción de la función de verosimilitud, de forma que ésta se puede dividir en dos partes diferenciadas. La primera, que se corresponde con la expresión (11) y la primera parte de la expresión (12), recoge la información de aquellas observaciones cuya duración es completa. El segundo miembro del producto de la ecuación (12) contempla la información que se deriva de las observaciones censuradas. Como se puede apreciar, esta aproximación permite incluir en la función de verosimilitud la única información de que se dispone acerca de las observaciones censuradas: su supervivencia hasta el momento en el que tiene lugar la censura.

Si tomamos logaritmos en la expresión (12), obtenemos la función de log verosimilitud[14]:

$$\ln L = \sum_{n=1}^{n} \{C_i \ln[f(t_i, x)] + (1-C_i) \ln[S(t_i, x)]\} \tag{13}$$

y, puesto que sabemos que $H(t) = -\ln [S(t)]$, y que $f(t) = h(t)\, S(t)$ podemos rescribir la expresión (13) únicamente en términos de la función de riesgo de la siguiente forma

$$\ln L = \sum_{n=1}^{n} \{C_i \ln[h(t_i, x)] - H(t_i, x)\} \tag{14}$$

[14] Uno de los supuestos fundamentales de construcción de la función de verosimilitud es que la censura tiene lugar de modo independiente. Un tipo de censura independiente es la aleatoria. Bajo este mecanismo, se asume que el momento de censura de un individuo es una variable aleatoria independiente entre individuos e independiente de los tiempos de supervivencia (censura de tipo I). Un ejemplo de censura aleatoria es la que tiene lugar cuando los individuos son seguidos hasta un momento del tiempo predeterminado e igual para todos. En general, la ida básica de la censura independiente es evitar que la causa por la que no observamos el suceso sea que el individuo tenga un riesgo muy alto o muy bajo de experimentar el suceso (Kalbfleisch y Prentice, 1980).

La ecuación (14) puede ser maximizada con el uso de métodos estándar. Como se ha comentado con anterioridad, para ello es necesario conocer $f(t_i, x)$, la función de densidad o, en términos de la expresión (14), es necesario especificar la función de riesgo. Dicha especificación puede tomar distintas formas según el supuesto que se realice acerca de cual es la variación de la función de riesgo conforme t se incrementa y cual sea el grado de parametrización impuesto y debería estar guiada por la teoría relevante sobre el tema objeto estudio. Dada la variedad de situaciones que puede sugerir la teoría, el análisis de supervivencia incluye gran número de modelos que tratan de adaptarse a los requerimientos de la misma.

De acuerdo con Kalbfleisch y Prentice (1980), podemos distinguir tres grupos de modelos en función de los supuestos que se realicen sobre la distribución de la función de riesgo: modelos de regresión exponencial y Weibull, modelos de tiempo de fallo acelerado y modelos de riesgo proporcional. Debe, en todo caso, quedar claro que los tres tipos de modelos están íntimamente relacionados. Los modelos de regresión exponencial y Weibull podrían considerarse como casos particulares, tanto del modelo de tiempo de fallo acelerado como del modelo de riesgo proporcional. El resto de esta sección está dedicada a analizar las características de los tres tipos de modelos señalados.

4.1. Modelos de regresión exponencial y Weibull

El primer grupo de modelos se basa en diferentes generalizaciones de la distribución exponencial. El modelo exponencial asume que la función de riesgo es constante, $h(t) = \lambda > 0$. Es decir, la distribución supone que las tasas de fallo son independientes de t y, por tanto, la probabilidad de fallo no depende del tiempo que el individuo lleva en el conjunto de riesgo (ausencia de memoria). Como señala Kiefer (1988), esta distribución es ampliamente utilizada para modelizar datos de duración, debido, principalmente, a su simplicidad de cálculo y estimación.

A partir del conocimiento de la función de riesgo y las relaciones entre los distintos tipos de funciones definidos en la sección anterior, podemos especificar las funciones de riesgo acumulado, supervivencia, densidad y distribución correspondientes al modelo exponencial de la siguiente forma (ver la figura 2 para una representación de las mismas):

$$H(t) = \lambda t \tag{15}$$

$$S(t) = \exp(-\lambda t) \tag{16}$$

$$f(t) = \lambda \exp(-\lambda t) \tag{17}$$

$$F(t) = 1 - \exp(-\lambda t) \tag{18}$$

El modelo exponencial puede modificarse fácilmente para tener en cuenta el efecto de diferentes variables explicativas (tanto cualitativas como cuan-

titativas) en los tiempos de supervivencia. Puesto que la función de riesgo ha de ser positiva, una forma de introducir dichas covariables en el modelo es utilizando la función exponencial[15]. Con esto, la función de riesgo del modelo exponencial es

$$h\,(t, x) = \lambda \exp(-x\beta) \tag{19}$$

donde x es el vector de covariables, λ es una constante y β representa el vector de parámetros[16]. Por tanto, un incremento en el valor de las variables explicativas tiene como consecuencia un incremento o disminución del riesgo de que un individuo experimente el suceso, según el valor del parámetro β asociado sea positivo o negativo, respectivamente.

Una de las generalizaciones del modelo exponencial da lugar al modelo de regresión Weibull. Este modelo considera el efecto de las covariables de modo similar, ya que la influencia de las variables explicativas en el riesgo es, de nuevo, multiplicativa. En este caso, la función de riesgo (covariables incluidas) queda especificada de la siguiente forma:

$$h\,(t, x) = \lambda p(\lambda t)^{p-1} \exp(x\beta) \tag{20}$$

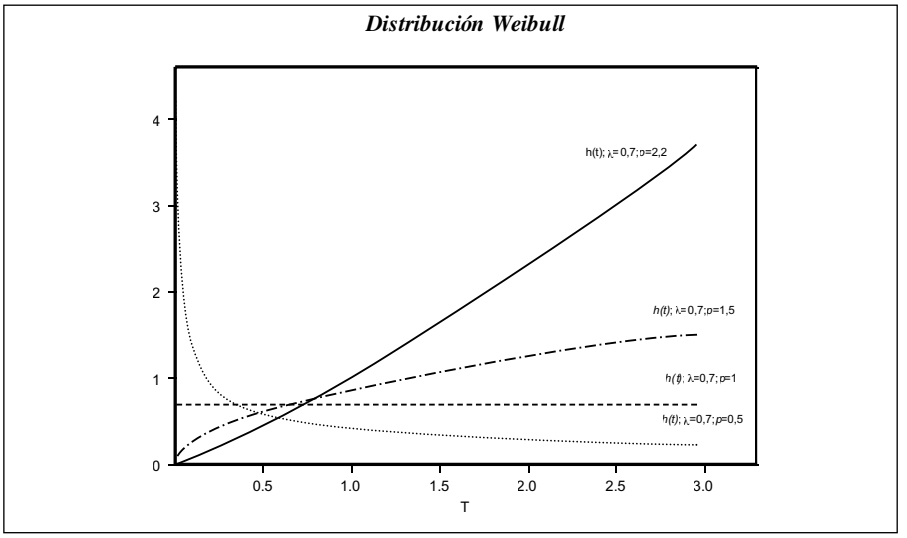
donde λ y p son constantes mayores que cero.

Como puede comprobarse fácilmente, la función de riesgo correspondiente al modelo Weibull queda reducida a la exponencial cuando el parámetro p es igual a 1. Es interesante observar como el modelo Weibull permite realizar distintos supuestos sobre la forma funcional de la función de riesgo, lo que le confiere gran flexibilidad. Según se aprecia en la Figura 3, este riesgo es monótono decreciente cuando $p < 1$ y monótono creciente cuando $p > 1$. Por tanto, la especificación de una distribución Weibull permite, a través de la estimación del parámetro , conocer si la función de riesgo es dependiente o no de la duración y si esa dependencia es positiva o negativa. De igual forma, el contraste nos indicaría si el modelo exponencial es o no apropiado para la muestra que poseemos y nos ayudaría a decidir si el modelo exponencial o el Weibull se adaptan mejor a los datos.

[15] Es importante destacar que el motivo por el que el modelo toma el nombre de exponencial es porque la variable aleatoria T se distribuye de acuerdo a una distribución exponencial y no por el hecho de que se use la función exponencial para introducir las covariables. La literatura también ha utilizado otras funciones con este propósito. Sin embargo, la ventaja de la función exponencial es que no es necesario restringir el valor del vector de parámetros para garantizar que el valor de su producto con las covariables sea mayor que cero, como requiere la función de riesgo (Kalbfleish y Prentice, 1980).

[16] Es importante destacar que, tal y como está expresado el modelo, el vector de parámetros, β, no incluye término constante. Una expresión alternativa del modelo vendría dada por $h(t; x) = \exp(x\beta)$, en la que el primer elemento del vector $x = (x_1,...,x_z)$ es igual a 1.

FIGURA 3.— *Distribución Weibull para distintos valores de p*



4.2. MODELOS DE TIEMPO DE FALLO ACELERADO

Una forma de rescribir los modelos exponencial y Weibull es a través de su analogía con la regresión lineal[17]. Si tomamos como ejemplo el caso del modelo Weibull, su expresión «logarítmico lineal» tomaría la siguiente forma:

$$\ln T = \alpha + x\gamma - \sigma\varepsilon \qquad (21)$$

donde $\alpha = -\ln\lambda$, $\gamma = -\sigma\beta$, $\sigma = p^{-1}$ y ε es el término del error, que se distribuye de acuerdo a la distribución del valor extremo[18]. A partir de esta expresión, es inmediato comprobar que el modelo exponencial, como caso particular del Weibull, también puede ser expresado en su versión log lineal[19].

Esta forma alternativa de representación ha originado un tipo de modelos que se conocen como modelos de tiempo de fallo acelerado (Kalbfleisch y

[17] La equivalencia entre las formas de tiempo de fallo acelerado y de riesgo proporcional del modelo Weibull y las relaciones entre sus parámetros puede verse en Cox y Oakes (1984: 71).

[18] La distribución del valor extremo recibe este nombre debido a que puede ser obtenida como distribución límite del valor más grande o más pequeño de n variables aleatorias independientes que, individualmente, tienen la misma distribución continua. Como puede intuirse, existe una relación matemática entre las distribuciones Weibull y del valor extremo. Si la variable aleatoria T se distribuye de acuerdo a una distribución Weibull, su logaritmo neperiano ($\ln T$) se distribuye conforme a la distribución del valor extremo. Es decir, en este caso la función de densidad del término del error ε es $\exp(\varepsilon - e^{\varepsilon})$ donde $-\infty < \varepsilon < +\infty$.

[19] Para ello, basta con saber que en el modelo exponencial el parámetro σ toma un valor igual a 1. De esta forma el modelo exponencial, en su forma log lineal, puede ser expresado como $\ln T = \alpha - x\beta + \varepsilon$.

Prentice, 1980) o modelos localización-escala (Lawless, 1982). Sin embargo, el conjunto de modelos incluidos dentro de esta categoría no queda restringido a los casos Weibull y exponencial. En general, si es el tiempo que transcurre hasta que ocurre el suceso (es decir, la duración) e $Y = \ln T$, estos modelos se pueden expresar como $Y = x\beta + \varepsilon$, donde ε es un término de error con una función de densidad dada, x es el vector de covariables y β el vector de parámetros a estimar. A partir de esta clase general de modelos de tiempo de fallo acelerado se han considerado diferentes casos particulares, según la distribución seguida por el término del error. Las más utilizadas suelen ser la normal, log-gamma, logística y las de valores extremos (Allison, 1984). Algunas de estas distribuciones, como la log-normal o la log-logística permiten obtener funciones de riesgo no monótonas.

Es importante destacar que en los modelos de tiempo de fallo acelerado la interpretación del vector de parámetros β es distinta a la de los modelos revisados en el apartado anterior (y también a los modelos de riesgo proporcional). Si bien en el caso de éstos últimos un parámetro con signo positivo indica un mayor riesgo de ocurrencia del suceso (una disminución de la duración), en los modelos de tiempo acelerado el efecto sería justamente el contrario: un incremento de la duración (y, por tanto, disminución del riesgo).

4.3. Modelos de riesgo proporcional

El último grupo de modelos que se distingue en la literatura pertenecen a lo que se han denominado modelos de riesgo proporcional (Cox, 1972). Su utilización es frecuente debido principalmente a su flexibilidad, dando lugar a multitud de aplicaciones empíricas y extensiones teóricas. Sea $h(t, x)$ la función de riesgo donde, de nuevo, t es un valor de T y x representa un vector de variables explicativas relativas al individuo. El modelo de riesgo proporcional se especifica del modo siguiente:

$$h(t, x) = \lambda_0(t) \exp(x\beta) \tag{22}$$

donde $\lambda_0(t)$ es la función de riesgo base[20]. Como se puede observar, el principal supuesto del modelo es que la función de riesgo de los individuos es un múltiplo de una función de riesgo arbitraria, no especificada y no negativa en el tiempo[21]. El modelo se denomina de riesgo proporcional debido a que,

[20] Por riesgo base se entiende aquel en el que incurre el individuo en ausencia de efectos de las covariables. Es decir, se trata del riesgo de que el suceso ocurra simplemente por el paso del tiempo, cuando el valor de todas las covariables es igual a cero.

[21] El supuesto de riesgo proporcional rara vez se cumple en la práctica (Vermunt 1997; Singer y Willett 1993) y, por tanto, una correcta aplicación del modelo exige comprobar si es razonable en el contexto en el que se está utilizando. Es importante destacar que mantener este supuesto cuando el riesgo es no proporcional puede resultar en estimadores sesgados y conclusiones erróneas al aplicar los test de significatividad (Kalbfleisch y Prentice 1980; Lagakos y Schoenfeld, 1984). Por ello la literatura ha propuesto distintas formas para comprobar esta hipótesis. Algunas de ellas pueden verse en Box-Steffensmeier y Zorn (2001).

dados dos individuos, el ratio entre sus tasas de riesgo se supone constante en cualquier momento del tiempo. Esto es, no existe interacción entre las covariables y el tiempo.

El rasgo más relevante de este tipo de modelos surge al considerar el procedimiento de estimación del vector de parámetros β y el grado de especificación de la función de riesgo base, $\lambda_0(t)$. Si bien los modelos (paramétricos) revisados hasta el momento son utilizados con cierta frecuencia para la modelización de la duración, éstos tienen ciertos inconvenientes (Allison, 1984). Tal vez el más importante es que para aplicarlos se necesita conocer el modo en que la función de riesgo depende del tiempo. Es decir, en términos del modelo de riesgo proporcional, es necesario concretar la forma de la función de riesgo base a partir de la información proporcionada por la teoría relevante. Por ejemplo, en el estudio de la difusión de innovaciones, la existencia de efectos epidémicos nos llevaría a pensar en una parametrización de la función de riesgo base que provocara una dependencia positiva de la duración que permitiera recogerlos. Con este objetivo, podríamos definir una función de riesgo base que se distribuyera con arreglo al modelo Weibull (en la que esperaríamos que el parámetro p tomara valores superiores a la unidad) y la estimación podría realizarse a través de los métodos de maximización habituales.

Desafortunadamente, esta información no está disponible en un gran número de situaciones, por lo que se han propuesto métodos de estimación alternativos. Uno de ellos es lo que Cox (1975) denomina el método de verosimilitud parcial. Este método de estimación elimina la función de riesgo base y, por tanto, reduce las exigencias de especificación respecto a los modelos paramétricos. Es importante destacar que este enfoque sólo tiene en cuenta el orden en el que se produce el evento y no el momento exacto de ocurrencia. Los estimadores obtenidos de este modo son consistentes y, asintóticamente, se distribuyen normalmente (Kalbfleisch y Prentice, 1980). No son completamente eficientes debido a la pérdida de información que se deriva de ignorar el tiempo exacto (Cox y Oakes, 1984). Sin embargo, esta pérdida de eficiencia es, con frecuencia, muy pequeña, incluso cuando la información utilizada procede del modelo paramétrico (Efron, 1977; Kalbfleisch, 1974; Oakes, 1977).

Por tanto, la principal ventaja del modelo de riesgo proporcional de Cox es que, frente a los modelos revisados con anterioridad, éste no exige especificar cuál es la dependencia de la función de riesgo sobre la duración. Ésta es la razón por la cual este modelo no sólo se considera más general (Allison, 1984), sino también más flexible que otras propuestas alternativas, por lo que su uso se ha difundido con rapidez.

Quizás uno de sus principales inconvenientes surge al aplicar el modelo sobre datos discretos. Si bien todos los modelos y funciones revisados en las secciones anteriores surgen en un contexto de tiempo continuo, en la práctica lo más frecuente es que la información sobre un determinado proceso y sus características asociadas esté disponible en períodos discretos (información diaria, mensual, anual o con otra periodicidad). Esto provoca que los modelos de duración continuos sean habitualmente aplicados sobre datos agrupados (en los que el tiempo es continuo, pero los sucesos son observados en intervalos discretos). Aunque esto puede suponer un problema menor en los mode-

los paramétricos, puede constituir grave inconveniente en el caso del modelo de Cox[22].

5. Extensiones de los modelos básicos

Como hemos visto, los modelos de duración básicos revisados en la sección anterior permiten valorar el impacto de las covariables sobre la duración (o, en otros términos, sobre el riesgo de finalización) de un determinado proceso. Sin embargo, en algunos casos el fenómeno estudiado es más complejo que la simple ocurrencia de un suceso de interés. Por ejemplo, podemos estar interesados en el estudio de un suceso que se repite para un mismo individuo, como puede ser la compra (y recompra) de una marca de un determinado producto. En otras ocasiones nos puede interesar explicar las distintas razones por las que el proceso llega a su fin. Una empresa puede desaparecer porque ha sido adquirida por otra, se ha fusionado o, simplemente, por haber quebrado. Es decir, podemos encontrarnos con razones que compiten a la hora de justificar la salida del individuo del conjunto de riesgo.

El interés por el estudio de sucesos con estructuras complejas y por una correcta especificación de los modelos de supervivencia ha llevado al desarrollo de extensiones que hacen extremadamente flexible su aplicación a distintos contextos. En esta sección se revisan algunas de estas extensiones. Concretamente, se hace referencia a la posibilidad de analizar sucesos repetidos o riesgos competidores, y se analiza el tratamiento que los modelos de duración ofrecen en presencia de heterogeneidad de distintos tipos.

5.1. Datos de duración con múltiples sucesos

En los modelos revisados hasta el momento sólo se contempla la posibilidad de que un sujeto experimente un único tipo de suceso y en una sola ocasión, lo que limita claramente su aplicabilidad en algunos contextos. Sin embargo, hay ocasiones en las que estaremos interesados en el análisis de un suceso que se repite en el tiempo. Este, por ejemplo, puede ser el caso de estudios en los que el evento de interés es la compra repetida de un producto por parte de un individuo o la duración de campañas publicitarias sobre un servicio. En estos ejemplos, no sólo nos encontramos con sucesos de un mismo tipo, sino que además la ocurrencia de los mismos presenta un orden claro en el sentido de que el segundo suceso no puede tener lugar hasta haber ocurri-

[22] Es importante destacar que, si los datos fueran realmente continuos, la posibilidad de empates no existiría y, con ello, el modelo de Cox sería aplicable sin necesidad de realizar ajuste alguno. Quizás las dos estrategias más utilizadas para resolver el problema son las derivadas de las propuestas de Breslow (1974) y Efron (1977), aunque existe, al menos, una tercera alternativa (menos eficiente en términos de cálculo que las aquí expuestas) que pasaría por construir la función de verosimilitud de forma «exacta» (Guide to Statistics, 1999).

do el primero (no puede haber recompra de un producto si no ha tenido lugar la primera compra). Sin embargo, en otras ocasiones, aunque los sucesos sean del mismo tipo, no va a existir una ordenación en su ocurrencia. Este sería el caso cuando estudiamos la entrada de una empresa en distintos mercados geográficos. Si definimos la entrada como suceso de interés, en principio no existiría ninguna razón para pensar que la introducción de una empresa en diferentes mercados deba seguir un orden concreto[23].

Aunque los ejemplos descritos no contemplan en su totalidad la complejidad de las situaciones que pueden presentarse, sí que ayudan a comprender la necesidad de una clasificación que nos permita decidir sobre el modelo a aplicar en una situación concreta. Una distinción que suele realizar la literatura es la que divide entre datos con sucesos repetidos y datos con riesgos competidores (Box-Steffensmeier y Jones, 1997; Box-Steffensmeier y Zorn, 1999)[24]. Se habla de *sucesos repetidos* cuando se trata de sucesos del mismo tipo en los que las observaciones se encuentran relacionadas. Este es el caso en los ejemplos del párrafo anterior. Tanto entradas en nuevos mercados, como compras o campañas publicitarias sucesivas, son sucesos idénticos cuya repetición está potencialmente relacionada. Tal y como hemos visto, estos sucesos pueden presentar un orden dado en su ocurrencia común para todos los individuos (primera compra, segunda compra...) o, por el contrario, tener un comportamiento totalmente aleatorio (entrada en mercados). Por otra parte, se habla de *riesgos competidores* cuando los sucesos son distintos y ocurren a observaciones no relacionadas. Este sería el caso si analizamos las distintas razones que llevan a la desaparición de una empresa (quiebra, adquisición, fusión...).

El principal problema que surge en el tratamiento estadístico de los *sucesos repetidos* es el de dependencia entre observaciones (Box-Steffensmeier y Zorn, 2001). En muchas aplicaciones, el supuesto de independencia entre las distintas repeticiones de un evento puede no ser realista, lo que afecta a la precisión de los estimadores. En estas situaciones, existen dos aproximaciones alternativas que permiten relajar dicha hipótesis (Cleves, 1999). La primera de ellas es la estimación de modelos de fragilidad (o modelos con heterogeneidad no observable), que se analizan posteriormente[25]. La segunda, en

[23] Dentro de la definición se suele incluir también el estudio de observaciones que, aún perteneciendo a distintos individuos, presentan una relación entre sí. En medicina, análisis de este tipo son comunes, por ejemplo, cuando se estudia el comportamiento de una enfermedad en miembros de una misma familia. A pesar de que los individuos poseen características que los hacen distintos a la hora de desarrollar una enfermedad, también comparten rasgos comunes (por ejemplo, la herencia genética) que explican la interrelación entre sus observaciones y la necesidad de su modelización.

[24] Therneau (1997) sugiere que las distinciones a realizar son: (1) si los sucesos tienen un orden y (2) si se trata de repeticiones de eventos del mismo tipo. La respuesta a estas dos cuestiones debería permitir seleccionar el tipo de modelo a aplicar.

[25] La denominación de modelos de fragilidad es utilizada con frecuencia en medicina para referirse a modelos en los que se contempla la posible existencia de heterogeneidad no observable. El origen del término surge de la posibilidad de que dicha heterogeneidad no observable pue-

la que nos centramos, consiste en la estimación de modelos de varianza corregida.

La idea general detrás de la aplicación de los modelos de varianza corregida parte de constatar que la matriz de covarianzas resultante de la aplicación de un modelo de duración tradicional sobre datos en los que existen observaciones dependientes es inapropiada para contrastar hipótesis (Struthers y Kalbfleisch, 1984; Lin y Wei, 1989; Box-Steffensmeier y Zorn, 2002). Por tanto, las técnicas disponibles suelen aplicar el modelo seleccionado a los datos como si las observaciones fueran independientes, para luego ajustar la matriz de covarianzas de modo que tenga en cuenta la posible correlación entre observaciones (Box-Steffensmeier y Zorn, 1999). El método es análogo a los utilizados en la regresión mínimo cuadrática ordinaria para tratar los problemas de heteroscedasticidad o autocorrelación.

Las extensiones a los modelos básicos para el caso de sucesos repetidos se han desarrollado fundamentalmente a partir del modelo de riesgo proporcional de Cox, aunque no existe ninguna razón para pensar que no puedan emplearse otras funciones de riesgo base (Cleves, 1999). Estas extensiones pueden dividirse en dos grupos, según se trate de sucesos ordenados o no ordenados. Quizás la situación más sencilla es aquella en la que los sucesos son del mismo tipo y no existe una ordenación en su ocurrencia (como era el caso de la entrada en distintos mercados geográficos), ya que la modelización sólo exigiría la modificación de la matriz de varianzas y covarianzas. Kelly y Lim (2000) realizan una revisión de los modelos disponibles cuando los sucesos están ordenados. Las extensiones más utilizadas son las propuestas por Andersen y Gill (1982), Wei, Lin y Weissfeld (1989) y Prentice, Williams y Peterson (1981). Las diferencias entre ellas descansan fundamentalmente en la forma de construir el conjunto de riesgo, las distintas formas de medir el tiempo hasta cada suceso y la posibilidad de que existan o no diferencias en las funciones de riesgo base entre las repeticiones de los eventos.

Aunque la casuística puede ser variada, quizás la situación más habitual de *riesgos competidores* en los estudios de economía de la empresa surge cuando existen razones mutuamente excluyentes por las que un individuo puede salir del conjunto de riesgo[26]. Si bien existen distintos enfoques para su tratamiento, la estrategia de modelización más simple y frecuente considera que los riesgos de sufrir cada uno de los sucesos son independientes entre sí. Bajo este supuesto de independencia, la función de verosimilitud puede ser dividida en factores en los cuales los riesgos distintos al de interés son trata-

da convertir a sujetos, por otra parte iguales, en más o menos frágiles ante una determinada enfermedad.

[26] Ejemplos de esto último son las distintas razones por las que puede terminar la vida de una empresa o el tiempo de permanencia de un trabajador. Sin embargo, puede haber situaciones donde los riesgos no sean mutuamente excluyentes y el sujeto bajo observación permanezca en riesgo de sufrir un suceso tras haber experimentado otro (u otros) de distinto tipo. En algunos casos, el período de riesgo de un sujeto para un suceso X puede continuar a pesar de que el suceso Y ha ocurrido y, sin embargo, la ocurrencia del suceso Y puede dar por finalizado el período de riesgo del suceso X (Yamaguchi, 1991).

dos como si sus correspondientes observaciones estuvieran censuradas (Kalbfleisch y Prentice, 1980; David y Moeschberger, 1978). Es decir, la estimación del modelo completo es equivalente a la estimación de modelos separados para cada tipo de riesgo en los que los riesgos distintos al de interés son tratados como observaciones censuradas (Zorn y Winckle, 2000).

5.2. HETEROGENEIDAD

Un último grupo de extensiones a los modelos básicos se ha desarrollado con el objetivo de hacer frente a la presencia de heterogeneidad de distintos tipos. En el contexto de los modelos de duración, la heterogeneidad surge cuando individuos homogéneos en las variables explicativas que son observables tienen distinto riesgo de sufrir el suceso objeto de estudio. Aunque las razones de la existencia de heterogeneidad pueden ser variadas, existen dos causas principales que pueden justificar su aparición (Box-Steffensmeier y Zorn, 1999): covariables no observables y observaciones que nunca experimentarán el suceso.

Una primera forma de heterogeneidad se debe a la existencia de *variables explicativas no observables*. Los problemas derivados de la presencia de heterogeneidad no observable han sido profusamente tratados en la literatura econométrica y son ampliamente conocidos por los investigadores en economía de la empresa. En el contexto del análisis de supervivencia, la consecuencia más conocida de la heterogeneidad no observable quizás sea la tendencia de la función de riesgo a mostrar dependencia negativa de la duración (Heckman y Singer, 1984). El motivo es que las observaciones pertenecientes a individuos con un mayor riesgo abandonan el conjunto de riesgo relativamente antes que aquéllos con un riesgo inferior, provocando la aparición de funciones de riesgo agregadas en las que éste se reduce con el tiempo (Petersen, 1995; Box-Steffensmeier y Zorn, 1999).

Las técnicas que se han propuesto para la solución de los problemas derivados de la heterogeneidad no observable en el contexto de los modelos de duración son análogas a las existentes para la regresión. Al igual que en ese caso, dos soluciones destacan frente a las demás (Petersen, 1995). La primera asume que la heterogeneidad no observable puede ser modelizada como un efecto aleatorio. Por tanto, este método presupone la existencia de covariables no observables que influyen sobre la función de riesgo y que se comportan de acuerdo a una función de distribución conocida. Aunque las funciones de distribución que se han utilizado con este fin son diversas, la elección más popular es la distribución gamma, en base, principalmente, a razones técnicas (Box-Steffensmeier y Zorn, 1999). El segundo método asume que la mencionada heterogeneidad es un efecto fijo. A pesar de su aparente atractivo, esta alternativa no se utiliza demasiado en los modelos de duración debido, fundamentalmente, a su desventaja relativa respecto a los modelos de efectos aleatorios (Andersen *et al.*, 1999).

La segunda forma de heterogeneidad aparece cuando la población objeto de estudio puede subdividirse entre los individuos que tarde o temprano expe-

rimentarán el suceso y aquellos que no lo sufrirán nunca (Box-Steffensmeier y Zorn, 1999). Los modelos de supervivencia tradicionales asumen la ocurrencia (observable o no) final del evento en algún momento del tiempo. Si bien este supuesto es razonable en estudios biomédicos, puede serlo menos cuando se trata de investigaciones en las áreas de marketing u organización de empresas. En algunos casos nos podremos encontrar con individuos que nunca compren un determinado producto por mucho que se prolongue el período de observación, o con empresas que nunca entren en un mercado.- La inclusión de este tipo de observaciones en el conjunto de riesgo puede provocar la sobreestimación de la función de supervivencia, con el consiguiente efecto sobre los parámetros a estimar (Price y Manatunga, 2001). Como consecuencia de ello, la literatura sobre análisis de supervivencia ha desarrollado modelos en los que se distingue entre individuos que experimentarán el suceso y los que nunca lo harán, denominados *modelos de curación* (Schmidt y Witte, 1988, 1989). La característica fundamental de estos modelos y su diferencia respecto a los revisados en la sección anterior es, precisamente, la división de la muestra en distintos grupos, según se trate de observaciones que «morirán» o no.

6. El análisis de supervivencia en economía de la empresa

Como se ha señalado en la introducción, si bien las aplicaciones del análisis de supervivencia son frecuentes en algunas disciplinas, su utilización es menos habitual en el ámbito del marketing o la organización de empresas. El objetivo de este apartado es ofrecer ejemplos de investigaciones en las que el análisis de supervivencia ha sido o puede ser aplicado con éxito con el ánimo de extender su uso allí donde es más adecuado que las técnicas tradicionales[27]. Con este fin, en primer lugar se revisan una serie de trabajos que han utilizado los modelos aquí expuestos para, posteriormente, ilustrar la aplicación de los mismos a través de dos ejemplos.

6.1. Modelos de duración en economía de la empresa

Quizás dos de las áreas en las que los modelos de duración se han utilizado con más frecuencia son los estudios sobre rotación de empleados y supervivencia de empresas. En el primer caso, y de modo análogo a las investigaciones sobre mercado de trabajo en economía (véase, por ejemplo, Lancaster, 1979) los artículos se han centrado en analizar los motivos por los que los empleados deciden permanecer o abandonar la organización en la que trabajan (Lee y Mowday, 1987). Aunque muchos de los estudios utilizan técnicas

[27] Es preciso notar que el objetivo de este apartado no es realizar una revisión exhaustiva de las áreas de investigación en las que el análisis de supervivencia ha sido o puede ser aplicado, sino proporcionar algunos ejemplos de temas en los que es posible su utilización.

distintas al análisis de supervivencia, algunos hacen uso de modelos de duración. Por ejemplo, Sheridan (1992) estudia la importancia de la cultura de la empresa sobre la retención de los empleados utilizando como variable dependiente el número de meses transcurridos desde la contratación hasta la fecha de salida de la organización. Más recientemente, Hom y Kinicki (2001) investigan el mecanismo a través del cual la insatisfacción en el puesto de trabajo motiva la rotación. Si bien en la mayor parte de los casos el modelo utilizado es el de riesgo proporcional de Cox, algunos autores se deciden por diseños más complejos. Trevor (2001) evalúa las interacciones entre las posibilidades de movilidad del trabajador y la satisfacción en el trabajo al predecir la rotación voluntaria. Este trabajo es, desde nuestro punto de vista, más interesante que los anteriores ya que ilustra la aplicación del modelo de Cox en el contexto de sucesos repetidos, haciendo uso de un modelo de varianza corregida como los analizados en la sección anterior.

Los modelos de duración también se han aplicado con relativa frecuencia dentro de la ecología de las poblaciones, uno de cuyos objetivos es el análisis de la supervivencia de las empresas (Hannan y Freeman, 1989). La utilización de los modelos revisados anteriormente en este contexto es una extensión natural de los estudios biomédicos, en los que se analiza el tiempo de supervivencia de una muestra de individuos sometidos a tratamiento. Dada la alta mortandad de las empresas en los primeros años de funcionamiento, no es extraño que los estudios se hayan preocupado, fundamentalmente, por el análisis de los factores que intervienen en el proceso. Este interés se ha extendido también a otras áreas, como la estrategia (Mata y Portugal, 2002) o la organización industrial (ver, por ejemplo, Mata y Portugal, 1995, o Audrestsch, Houweling y Thurik, 2000). Segarra y Callejón (2002) utilizan el modelo de riesgo proporcional de Cox para examinar la supervivencia de empresas españolas a partir de la base de datos del Directorio Central de Empresas. De nuevo, una extensión interesante (a un contexto de datos agrupados) es la que aplican Mata y Portugal (2002) para analizar las diferencias en las tasas de mortandad entre empresas portuguesas y empresas de capital extranjero.

Sin embargo, la analogía con los estudios biomédicos no queda limitada al contexto de la supervivencia de empresas y puede ser extendida, entre otros ámbitos, al estudio del mantenimiento de relaciones o contratos entre empresas. Éste es el caso de investigaciones que utilizan distintas versiones de los modelos revisados en el apartado anterior para identificar los factores de éxito o fracaso de las alianzas. Por ejemplo, Barkema *et al.* (1997) utilizan un modelo de tiempo de fallo acelerado en el que se asume una distribución Weibull para analizar la influencia del aprendizaje sobre la supervivencia de las alianzas llevadas a cabo por 25 empresas holandesas en su proceso de expansión. Del mismo modo, Park y Ungson (1997) utilizan un modelo de tiempo de fallo acelerado con distribución logarítmico normal con un objetivo similar.

Además de las áreas mencionadas, en las que la aplicación de los modelos de supervivencia surge como una extensión más o menos natural, existen otras con gran potencial para su utilización, como la difusión de innovaciones, el estudio de la entrada de empresas en nuevos mercados o los análisis de fidelidad de clientes. El estudio de la difusión de innovaciones es un área donde

las técnicas de análisis de supervivencia se han aplicado de forma parcial. Si bien los modelos de duración se han utilizado para el estudio del tiempo hasta la primera adopción de una nueva tecnología (Hannan y McDowell, 1984; Sharma, 1993; Karshenas y Stoneman, 1993; Baptista, 2000), no existe ningún intento conocido de incorporar su uso al análisis de la velocidad con la que la difusión tiene lugar dentro de las empresas. La utilización de los modelos de duración en este caso parece una extensión lógica de su aplicación a los estudios del tiempo hasta la primera adopción. De hecho, Karshenas y Stoneman (1995) sugieren que si se definen distintos estados en el proceso de adopción interno de una tecnología, los modelos de duración pueden utilizarse para estimar el tiempo hasta un determinado momento de adopción.

De forma similar, una parte importante de la literatura de entrada ha dedicado sus esfuerzos al estudio de los factores que explican el momento en el que ésta se produce. Sin embargo, con frecuencia y salvo algunas excepciones (Mitchell, 1989; Baum y Korn, 1996), los estudios han renunciado a la posibilidad de integrar en el análisis la decisión de entrar o no hacerlo y el momento de la misma (Fuentelsaz *et al.*, 2002). Como ya se ha comentado, las técnicas tradicionalmente utilizadas han sido modelos de variable dependiente limitada (logit o probit en sus distintas versiones) que han impedido el aprovechamiento adecuado de la información contenida en la dimensión longitudinal de los datos. En este sentido, los modelos de supervivencia se consideran herramientas más adecuadas para el estudio del orden de entrada en los mercados (Lieberman y Montgomery, 1998). Siguiendo este consejo, Fuentelsaz *et al.* (2002) aplican la extensión del modelo de riesgo proporcional de Cox propuesta por Andersen y Gill (1982) al estudio del momento de entrada de las cajas de ahorros españolas en los distintos mercados provinciales[28].

Es importante destacar que la entrada (o salida) en un mercado no es la única acción empresarial susceptible de ser analizada a través de la aplicación de modelos de duración, sino que el análisis es ampliable al estudio de otras decisiones como el lanzamiento de nuevos productos o el inicio de una campaña publicitaria. La literatura sobre dinámica competitiva entre empresas enfatiza, precisamente, la importancia de la evaluación de las acciones competitivas a la hora de realizar valoraciones sobre la rivalidad (Chen, 1996). Es de esperar que el tipo, la velocidad, la frecuencia y la intensidad de las acciones competitivas de las empresas sean diferentes dependiendo del nivel de rivalidad entre ellas. En este contexto, los modelos de supervivencia ofrecen una nueva herramienta con la que poder analizar parte de la información contenida en dichas acciones que sirva de complemento a evaluaciones basadas única y exclusivamente en medidas de resultados.

Por último, y como ya se ha resaltado anteriormente, los modelos de duración también se han aplicado al estudio de la fidelidad del consumidor. Los proyectos de retención de consumidores estudian las características de aque-

[28] Baum y Korn (1996) y Haveman y Nonnemaker (2000) también utilizan modelos de duración para analizar entrada, salidas y expansión de empresas al analizar los efectos del contacto multimercado.

llos clientes que rompen más rápidamente la relación establecida con la empresa para poder diseñar planes de actuación encaminados a la retención de los mismos (Shaomin, 1995). El análisis de supervivencia puede ser, por tanto, utilizado en este contexto con el objetivo de determinar los factores que afectan a la duración de la relación empresa-consumidor y que ayuden a clasificar a estos últimos a partir del riesgo de ruptura de dicha relación[29].

6.2. Aplicaciones del análisis de supervivencia[30]

En esta sección se presentan dos ejemplos que ilustran la aplicación de algunos de los modelos que han sido expuestos con anterioridad[31]. El primero es un caso ficticio a partir del que se muestra la utilización de los modelos de regresión Weibull y de tiempo de fallo acelerado y la conexión existente entre ellos. El segundo presenta un análisis de la entrada de las cajas de ahorro españolas en nuevos mercados geográficos tras la liberalización del sector. Para ello, se utiliza la versión de Andersen y Gill (1982) del modelo de riesgo proporcional de Cox (1972), así como la extensión de dicho modelo al contexto de sucesos repetidos.

Ejemplo ficticio: difusión de una nueva tecnología

Los modelos propuestos en el análisis de supervivencia pueden utilizarse para estudiar el tiempo de difusión de una nueva tecnología dentro de una empresa. El ejemplo elegido en este caso ilustra la aplicación del modelo Weibull y muestra la conexión existente entre su interpretación en términos de riesgo relativo y la resultante de su transformación a modelo de tiempo de fallo acelerado. Como se señaló al exponer los modelos básicos, la relación entre ambos se manifiesta de forma sencilla sobre todo a través de los parámetros β y p del modelo Weibull, de forma que, a partir de su conocimiento, podemos calcular sus correspondientes del modelo de tiempo acelerado $\gamma = -\sigma\beta$ y $\sigma = p^{-1}$.

En este caso, la variable dependiente elegida es el tiempo, medido en años, transcurrido desde que la empresa adopta la nueva tecnología hasta que fina-

[29] Los modelos para el análisis de sucesos repetidos ofrecen la posibilidad de estudiar no sólo la duración de la relación empresa-consumidor, sino también de obtener información más detallada de su comportamiento a través del examen de la repetición de la compra.

[30] Los ejemplos incluidos en esta sección únicamente pretenden ilustrar la utilización de algunos de los modelos presentados en el trabajo y nunca sacar conclusiones válidas para las literaturas en las que se incluyen.

[31] Las estimaciones que se presentan en esta sección han sido realizadas con los programas Stata (Stata Corp. 2001) y SPLUS (Guide to Statistics: S-Plus 2000) (versiones 7.0 y 2000 respectivamente). Ambos ofrecen un amplio rango de posibilidades para estimar distintos modelos de duración, entre los que se encuentran los expuestos en este trabajo. Sin embargo, existen otros muchos programas estadísticos que incluyen módulos destinados a estimar modelos de supervivencia que implican mayor o menor grado de sofisticación. Algunos ejemplos son SPSS, SAS, LIMDEP, BMDP o GAUSS.

liza el proceso de difusión interno. Para un caso simple como éste, en el que no aparecen covariables cuyos valores cambian durante el período de observación, la construcción de la base de datos para proceder a la estimación del modelo es relativamente sencilla. El cuadro 1 muestra dicha construcción e incluye, además de la identificación de la propia empresa, los valores de dos de las variables sugeridas por la literatura (el retraso en la adopción y el tamaño de la empresa), así como otras dos variables especialmente relevantes en este tipo de modelos. La primera es el indicador de censura, que toma el valor uno si el proceso de difusión ha finalizado (hemos observado el suceso de interés) y cero si, por el contrario, todavía no lo ha hecho. La última columna de la tabla muestra el tiempo transcurrido hasta la finalización del proceso de difusión o hasta el momento de observación. Por tanto, la lectura conjunta de las dos últimas columnas para la «Empresa 9» nos permite concluir que el período de difusión se ha observado por completo («Censura=1») y su duración ha sido de 6 años («Tiempo=6»). Por el contrario, para la «Empresa 12» la observación del período de difusión fue interrumpida en el séptimo año («Censura=0» y «Tiempo=7»).

CUADRO 1.—*Construcción de la base de datos para el modelo Weibull (covariables invariantes)*

Empresa	Retraso en la adopción (años)	Tamaño de la empresa (millones de euros)	Censura	Tiempo (años)
…	…	…	…	…
9	8	0,358	1	6
10	7	0,368	0	6
11	7	0,378	1	6
12	7	0,388	0	7
13	6	0,398	1	7
…	…	…	…	…

El cuadro 2 muestra los resultados de una hipotética estimación de este modelo. La segunda columna del cuadro muestra los resultados de la estimación en su forma Weibull, mientras que la tercera presenta la parametrización alternativa en la forma de modelo de tiempo de fallo acelerado[32]. Se puede comprobar como los coeficientes de ambas columnas son equivalentes a partir de la utilización de las expresiones anteriores.

Es importante destacar la diferencia en la interpretación existente entre ambos modelos a partir de las estimaciones que se presentan. En el modelo Weibull un signo positivo implica un incremento en la tasa de riesgo y un signo negativo una disminución de la misma. Así, el signo positivo del retraso en la adopción implica un incremento en la probabilidad instantánea de que el proceso de difusión termine (dado que no lo ha hecho hasta el momento). Por el contrario, en el caso del modelo de tiempo acelerado el efecto de las variables es el de «alargar» o «reducir» el tiempo transcurrido hasta el final del pro-

[32] Los resultados han sido calculados a partir de los comandos «stset» y «streg» de Stata .

ceso de difusión. Para esa misma variable, la interpretación del signo del coeficiente en el modelo de tiempo de fallo acelerado nos llevaría a la conclusión de que su efecto es el de «acortar» el tiempo de difusión.

CUADRO 2.—*Estimación de un modelo de difusión ficticio*

Variables explicativas	Weibull (β)	Tiempo de fallo acelerado ($\gamma = -\sigma\beta$)
Constante	–18,888*** (–5,62)	2,045*** (7,24)
Retraso en la adopción	0,550*** (2,69)	–0,059*** (–2,68)
Tamaño de la empresa	–11,036*** (–3,19)	1,195*** (3,48)
Parámetros	p = 9,233 (1,116)	$\sigma = p^{-1}$ =0,108 (0,130)
Ratio de verosimilitud	82,42***	82,42***
Número de observaciones	50	50

*** Coeficientes estadísticamente significativos al 1%. T-ratios entre paréntesis, excepto en el caso de los parámetros σ y p y para los que se presentan los errores estándar.

Finalmente, la magnitud de los coeficientes es fácilmente interpretable a partir del ratio entre las tasas de riesgo de dos individuos, $h(t, x_1) / h(t, x_2) = \exp(x_1 - x_2)\beta]$. Así, para un incremento unitario en la variable *retraso en la adopción* la tasa de riesgo se multiplica por 1,733 (exp[0,550•1]). En el caso del parámetro p su valor, mayor que uno, nos indica que la tasa de riesgo es creciente en el tiempo. Es interesante señalar como dicho parámetro puede ayudar a comprobar si el modelo Weibull elegido es el adecuado frente a la especificación exponencial. La decisión puede tomarse con la aplicación del contraste $p = 1$ a partir de la expresión $z = (p–1) / e(p)$, donde $e(p)$ es el error estándar correspondiente al parámetro p (Box-Steffensmeier y Jones, 1997).

Aplicación del modelo de Cox al análisis de la entrada

Otra de las posibles aplicaciones de los modelos de duración es el estudio del proceso de entrada de empresas en nuevos mercados. En este caso, queremos responder a algunas de las cuestiones a las que ya hemos hecho referencia en la sección 2: ¿son diferentes los entrantes tempranos y los entrantes tardíos? O, planteado de otro modo: ¿Qué características tienen las empresas pioneras? La base de datos utilizada para dar respuesta a estas preguntas describe el comportamiento diversificador de las cajas de ahorros españolas entre los años 1986 y 1999. Por tanto, el suceso de interés es la entrada de una caja de ahorros en una determinada provincia española[33].

[33] Este análisis está basado en Fuentelsaz, Gómez y Polo (2002).

El modelo que utilizamos es el de riesgo proporcional propuesto por Cox (1972). Como ya se ha señalado, se trata de un modelo flexible que es adecuado para aquellas situaciones en las que no existe una hipótesis sobre el tipo de dependencia de la función de riesgo sobre la duración y permite utilizar covariables cuyos valores cambian con el tiempo. Dado que cada entidad puede realizar su entrada en distintos mercados geográficos (provincias) y, a priori, no existe una ordenación entre dichas entradas, utilizamos la extensión del modelo desarrollada por Andersen y Gill (1982)[34].

El cuadro 3 muestra los resultados de la aplicación de dicho modelo al estudio de la entrada[35]. Como podemos observar, el ajuste global es bueno puesto que la hipótesis de que los coeficientes del modelo sean todos igual a cero es rechazada al 99% (ratio de verosimilitud estadísticamente significativo). Los signos de los coeficientes que se presentan en la columna (1) nos muestran que el tamaño de la entidad, la posición en el mercado interbancario y la proximidad del mercado objetivo tanto al mercado original como a la expansión realizada son variables que reducen el tiempo transcurrido desde la eliminación de las barreras legales a la apertura de oficinas hasta que se produce la entrada. Por el contrario, el incremento de la concentración en el mercado original y la competencia potencial en el mercado objetivo hacen dicha duración más larga.

La última columna del cuadro 3 muestra el cambio porcentual en la tasa de riesgo para cada una de las variables que resultan ser significativas. Para una variable independiente continua dicho valor se calcula a partir de la siguiente expresión:

$$100[\exp(\beta_j (x+ \theta)) - \exp(\beta_j x)] / \exp[\beta_j x]$$

que proporciona el cambio porcentual en la tasa de riesgo para una variación unitaria en la variable independiente x. De este modo, el efecto de una variación en una desviación típica (1,24) de la variable que mide el tamaño de la entidad supone un incremento en la tasa de riesgo del 346%. Por el contrario, si la concentración en el mercado original de la entidad se incrementa en su desviación típica (0,04) la tasa de riesgo disminuye en un 43%.

Los cálculos a realizar son similares para el caso de una variable ficticia. Así, la proximidad del mercado objetivo a las provincias en las que inicialmente operaba la entidad incrementa el riesgo de entrada en un 1.015% (100(exp(2,412*1)-exp(2,412*0))/exp(2,412*0)). Por el contrario, si esa proximidad es a mercados hacia los que se ha realizado la expansión, el incremento es del 570%.

Como ya se ha señalado, uno de los problemas que surge en el tratamiento estadístico de los sucesos repetidos es la dependencia de las observa-

[34] La construcción de la base de datos es más compleja que en el caso básico y difiere según el tipo de modelo utilizado. Para una descripción de su construcción puede consultarse Therneau y Hamilton (1997).

[35] El comando utilizado en este caso es «coxph» de SPLUS (Guide to Statistics, 1999).

ciones. En nuestro caso, parece razonable pensar que dicha dependencia puede aparecer por la coordinación de actividades realizada desde la dirección. De este modo la expansión tendría lugar no como resultado de decisiones de entrada independientes sino como consecuencia de un plan de expansión global de las entidades, en cuyo caso las entradas de una entidad en distintas provincias estarían correlacionadas, con el consiguiente efecto sobre la precisión de los estimadores. Para controlar por esa posibilidad, la columna 2 corrige la matriz de covarianzas tal como se ha explicado en la sección quinta. Como se puede observar, el efecto de la corrección sólo tiene consecuencias sobre los errores estándar y los estadísticos de significación asociados a los estimadores, que disminuyen en valor absoluto en la mayor parte de los casos.

CUADRO 3.—*Factores que afectan al momento de entrada de las cajas de ahorro españolas*

Variables explicativas	(1) Coeficiente	(2) Coeficiente	Cambio porcentual en la tasa de riesgo
Tamaño de la entidad (logaritmo)	1,207*** (9,778)	1,207*** (6,110)	346
Posición en el mercado interbancario	3.412* (2.387)	3.412* (1,804)	31
Beneficios de la entidad	0,025 (0,175)	0,025 (0,141)	—
Proximidad al mercado original	2,412*** (12.281)	2,412*** (4,282)	1.015
Proximidad a la expansion	1,903*** (8,015)	1,903*** (4,120)	570
Número de mercados en los que opera	0,054 (1,515)	0,054 (1,388)	—
Número de mercados en los que opera (componente cuadrático)	–0,001* (–1,653)	–0,001 (–1,500)	—
Concentración (mercado original)	–14,220*** (–4,873)	–14,220*** (–3,317)	–43
Concentración (mercado objetivo)	–2,543 (–1,406)	–2,543 (–1,056)	—
Competencia potencial en el mercado objetivo	–0,091*** (–3,430)	–0,091*** (–2,849)	–28
Depósitos por habitante	0,069 (0,144)	0,069 (0,080)	—
Densidad de población	–0,200 (–0,333)	–0,200 (–0,279)	—
Crecimiento del Mercado	0,637 (0,258)	0,637 (0,297)	—
Test del ratio de verosimilitud	853***	853***	—
Número de observaciones	34.529	34.529	—

***, **, * Coeficientes estadísticamente significativos al 1%, 5% y 10%, respectivamente. t-ratios entre paréntesis.

7. Conclusión

El objetivo de este artículo ha sido ofrecer los fundamentos del análisis de supervivencia y sus posibilidades de aplicación a distintos ámbitos dentro de la investigación en economía de la empresa. Como se ha comentado, la utilización de esta técnica es relativamente común en áreas como la biomedicina o la ingeniería. Sin embargo, su uso en disciplinas como el marketing o la organización de empresas es más bien escasa. Por tanto, nuestro propósito ha sido ofrecer una descripción de los modelos disponibles e ilustrar su aplicación a partir de la revisión de algunos artículos que han utilizado una u otra versión de los mismos con el ánimo de facilitar y promover su uso en aquellos contextos en los que éstos son apropiados.

Es preciso destacar que la revisión realizada en este trabajo es, necesariamente, limitada. Dadas las características de las investigaciones en economía de la empresa, el énfasis se ha puesto en la descripción de las técnicas paramétricas y semiparamétricas, obviando las aproximaciones no paramétricas que existen en la literatura por su menor relevancia. De igual modo, los modelos presentados han sido diseñados para el análisis de eventos en un contexto de tiempo continuo. No se ha hecho referencia, por tanto, a los modelos desarrollado para el análisis de sucesos cuya ocurrencia sólo puede tener lugar en momentos discretos del tiempo o cuando los datos sólo están disponibles de forma agrupada[36]. El tratamiento de ambos temas (modelos no paramétricos y para datos discretos) puede ser consultado en los libros de Kalbfleisch y Prentice (1980), Allison (1984), Cox y Oakes (1984) o Yamaguchi (1991)[37].

Referencias bibliográficas

ALLISON, P. D. (1984), *Event History Analysis*, Sage University Paper series on Quantitative Applications in the Social Sciences, series núm. 07-001, Beverly Hills y Londres, Sage.

ANDERSEN, P. K. y Gill, R. D. (1982), «Cox's regression model for counting processes: a large sample study», *Annals of Statistics*, 10, 1100-1120.

ANDERSEN, P. K.; KLEIN, J. P. y ZHANG, M. J. (1999), «Testing for centre effects in multi-centre survival studies: A Monte Carlo comparison of fixed and random effects tests», *Statistics in Medicine,* 18, 1489-1500.

ASPLUND, M. y SANDIN, R. (1999), «Survival of new products», *Review of Industrial Organization*, 15, 219-237.

AUDRESTSCH, D. B.; HOUWELING, P. y THURIK, A. R. (2000), «Firm Survival in the Netherlands», *Review of Industrial Organization*, 16, 1-11.

[36] Aunque en la mayor parte de los casos los datos de que se disponen aconsejan la utilización de modelos de tiempo discreto, éstos son poco utilizados. Una de las razones que explican este hecho es que ambas aproximaciones suelen ofrecer resultados similares.

[37] Los libros de Allison (1984) y Yamaguchi (1991) ofrecen una introducción a las técnicas de análisis de supervivencia más accesible para el lector no especializado que los de Kalbfleisch y Prentice (1980) o Cox y Oakes (1984).

BAIN, J. S. (1956), *Barriers to new competition: their character and consequences in manufacturing industries,* Cambridge, MA., Harvard University Press.

BAPTISTA, R. (2000), «Do innovations diffuse faster within geographical clusters?», *International Journal of Industrial Organisation*, 18, 515-535.

BARKEMA, H. G.; SHENKAR, O.; VERMEULEN, F. y BELL, J. H. J. (1997), «Working Abroad, Working with Others: How Firms Learn to Operate International Joint Ventures», *Academy of Management Journal*, 40(2), 426-442.

BAUM, J. A. C. y KORN, H. J. (1996), «Competitive Dynamics of Interfirm Rivalry», *Academy of Management Journal*, 39(2), 255-291.

BECK, Nathaniel (1998), «Modeling Space and Time: The Event History Approach», en Elinor Scarbrough y Eric Tanenbaum (eds.), *Research Strategies in the Social Sciences*, Oxford, Oxford University Press.

BOX-STEFFENSMEIER, J. M. y JONES, B. S. (1997), «Time Is of the Essence: Event History Models in Political Science», *American Journal of Political Science*, vol. 41(4), págs. 1414-1461.

BOX-STEFFENSMEIER, J. M. y ZORN, C. (1999), «Modeling Heterogeneity in Duration Models», *Annual Meeting of the Political Methodology Society*, Texas, College Station.

— (2001), «Duration Models and Proportional Hazards in Political Science», *American Journal of Political Science,* 45(octubre), 972-88.

— (2002), «Duration Models for Repeated Events», *Journal of Politics,* en prensa.

BRESLOW, N. E. (1974), «Covariance analysis of censored survival data», *Biometrics*, 30, 89-99.

CHEN, M. J. (1996), «Competitor analysis and interfirm rivalry: toward a theoretical integration», *Academy of Management Review*, 21(1), 100-134.

CHUNG, C.; SCHMIDT, P. y WITTE, A. D. (1991), «Survival Analysis: A Survey», *Journal of Quantitative Criminology,* vol. 7, págs. 59-98.

CLEVES, M. (1999), «Analysis of Multiple Failure-Time Data with Stata», *Stata Technical Bulletin*, 49, 30-39.

COTERILL, R. W. y HALLER, L. E. (1992), «Barrier and queue effects: a study of leading US supermarket chain entry patterns», *The Journal of Industrial Economics*, XL (4), 427-440.

COX, D. R. (1972), «Regression Models and Life-Tables (with discussion)», *Journal of the Royal Statistical Society,* 34 (serie B), 187-202.

— (1975), «Partial Likelihood», *Biometrika,* 62, 269-276.

COX, D. R. y OAKES, D. (1984), *Analysis of survival data,* Londres y Nueva York, Chapman & Hall.

CYERT, R. M. y MARCH, J. G. (1963), *A behavioral theory of the firm*, Englewood Cliffs, NY, Prentice Hall.

DAVID, H. A. y MOESCHBERGER, M. (1978), *Theory of Competing Risks*, Londres, Griffin.

DIERMEIER, D. y STEVENSON, T. (1999), «Cabinet Survival and Competing Risks», *American Journal of Political Science*, 43(4), 1051-1068.

EFRON, B. (1977), «The efficiency of Cox's likelihood function for censored data», *Journal of the American Statistical Association*, 72, pág. 557-565.

FLAVIÁN, C.; GÓMEZ, J.; MARTÍNEZ, E. y POLO, Y. (1998), «Factores determinantes del nivel de empleo en el sector detallista. Una aplicación del análisis de supervivencia», *Revista Española de Investigación de Marketing ESIC*, 2 (2), 7-25.

FLEMING, T. y HARRINGTON, D. (1991), *Counting Processes and Survival Analysis,* Nueva York, Wiley.

FUENTELSAZ, L. y GÓMEZ, J. (2001), «Strategic and Queue Effects in Spanish Banking», *Journal of Economics and Management Strategy*, 10 (4), 529-563.

FUENTELSAZ, L.; GÓMEZ, J. y POLO, Y. (2002), «Followers entry timing: Evidence from

the Spanish banking sector after deregulation», *Strategic Management Journal*, 23 (3), 245-264.

Guide to Statistics: S-Plus 2000 Modern Statistics and Advanced Graphics (1999), Seattle, WA: MathSoft Data Analysis Products Division.

HANNAN, M. T. y FREEMAN, J. (1989), *Organizational Ecology*, Cambridge, MA., Havard University Press.

HANNAN, T. H. y MCDOWELL, J. M. (1984), «The Determinants of Technology Adoption: The Case of the Banking Firm», *Rand Journal of Economics*, 15, (3), 328-335.

HAVEMAN, H. A. (1993), «Organizational size and change: Diversification in the savings and loan industry after deregulation», *Administrative Science Quarterly*, 38, 20-50.

HAVEMAN, H. A. y NONNEMAKER, Lynn (2000), «Competition in Múltiple Geographic Markets: The Impact on Growth and Market Entry», *Administrative Science Quarterly*, 45, 232-267.

HECKMAN, J. J. y SINGER, B. (1984), «Econometric Duration Analysis», *Journal of Econometrics*, 24, 63-132.

HOM, P. W. y KINICKI, A. J. (2001), «Toward a Greater Understanding of how Dissatisfaction Drives Employee Turnover», *Academy of Management Journal*, 44(5), 975-987.

HOVERSTAD, R.; MONCRIEF, W. C. III y LUCAS, G. H. Jr. (1990), «The Use of Survival Analysis to Examine Sales Force Turnover of Part-Time and Full-Time Sales Employees», *International Journal of Research in Marketing*, vol. 7(2), páginas 109-120.

KALBFLEISCH, J. D. (1974), «Some Efficiency Calculations for Survival Distributions», *Biometrika,* 61, 31-28.

KALBFLEISCH, J. D. y PRENTICE, R. L. (1980), *The Statistical analysis of failure time data*, Nueva York, John Wiley.

KARSHENAS, M. y STONEMAN, P. L. (1993), «Rank, stock, order, and epidemic effects in the difusión of new process technologies: an empirical model», *Rand Journal of Economics,* vol. 24(4), págs. 503-528.

KARSHENAS, M. y STONEMAN, P. (1995), «Technological diffusion», en P. Stoneman (ed.), *Handbook of the Economics of Innovation and New Technology,* Blackwell.

KELLY, P. J. y LIM, L. L. Y. (2000), «Survival analysis for recurrent event data: an application to childhood infectious diseases», *Statistics in Medicine*, 19, 13-33.

KIEFER, N. M. (1988), «Economic Duration Data and Hazard Functions», *Journal of Economic Literature,* vol. 26, págs. 646-679.

LAGAKOS, S. W. y SCHOENFELD, D. A. (1984), «Properties of Proportional-Hazards Score Tests Under Misspecified Regression Models», *Biometrics,* 40, 1037-1048.

LANCASTER, T. (1979), «Econometric Methods for the Duration of Unemployment», *Econometrica*, 47(4), 939-956.

— (1990), *The econometric analysis of transition data,* Cambridge, Cambridge University Press.

LAWLESS, F. J. (1982), *Statistical models and methods for lifetime data*, Nueva York, Wiley.

LEE, T. W. y MOWDAY, R. T. (1987), «Voluntarily Leaving and Organization: An Empirical Investigation of Steers and Mowday's Model of Turnover», *Academy of Management Journal*, 30(4), 721-743.

LIAO, T. T. (1994), *Interpreting Probability Models: Logit, Probit, and Other Generalised Linear Models,* Sage University papers on Quantitative Applications in the Social Sciences, 07-101, Thousand Oaks, CA.

LIEBERMAN, M. B. y MONTGOMERY, D. B. (1998), «First-mover (dis)advantages: retrospective and link with the resource-based view», *Strategic Management Journal*, 19 (12), 1111-1125.

LIN, D. Y. y WEI, L. J. (1989), «The Robust Inference for the Cox Proportional Hazards Model», *Journal of the American Statistical Asoociation*, 84 (408), 1074-1078.

LUOMA, M. y LAITINEN, E. K. (1991), «Survival Analysis as a Tool for Company Failure Prediction», *Omega,* Oxford, 1991, vol. 19(6), págs. 673-689.

MATA, J. y PORTUGAL, P. (1995), «Life Duration of New Firms», *Journal of Industrial Economics*, 42, 227-245.

— (2002), «The Survival of New Domestic and Foreign-Owned Firms», *Strategic Management Journal*, 23, 323-343.

MITCHELL, W. (1989), «Whether and whem? Probability an timing of incumbents' entry into emerging industrial subfields», *Administrative Science Quarterly*, 34, 208-230.

MORITA, J. G.; LEE, T. W. y MOWDAY, R. T. (1993), «The regression analog to Survival Analysis: A Selected Application to Turnover Research», *Academy of Management Journal,* vol, 36 (6), págs. 1430-1464.

OAKES, D. (1977), «The Asymptotic Information in Censored Survival Data», *Biometrika*, 59, 441-448.

PARK, S. H. y UNGSON, G. R. (1997), «The Effect of National Culture, Organizational Complementariry and Economic Motivation on Joint Venture Dissolution», *Academy of Management Journal*, 40(2), 279-307.

PETERSEN, T. (1985), «Fitting Parametric Survival Models with Time-Dependent Covariates», *Applied Statistics*, 35, 281-288.

— (1995), «Analysis of Event Histories», *Handbook of Statistical Modeling for the Social and Behavioral Sciences,* Nueva York, Plenum Press, págs. 453-517.

PRENTICE, R. L.; WILLIAMS, B. J. y PETERSON, A. V. (1981), «On the Regression Analysis of Multivariate Failure Time Data», *Biometrica*, 68, 373-379.

PRICE, D. L. y MANATUNGA, A. K. (2001), «Modelling survival data with a cured fraction using frailty models», *Statistics in Medicine*, 20, 1515-1527.

SCHMIDT, P. y WITTE, A. D. (1988), *Predicting Recidivism Using Survival Models*, Nueva York, Springer-Verlag.

SCHMIDT, P. y WITTE, A. D. (1989), «Predicting Recidivism Using "Split-Population" Survival Time Models», *Journal of Econometrics*, 40(1), 141-159.

SHAOMIN, L. (1995), «Survival Analysis», *Marketing Research*, vol. 7(4), 17-23.

SHARMA, S. (1993), «Behind the Diffusion Curve: An Analysis of ATM adoption», *Working Paper, núm. 686,* Los Ángeles, University of California, Department of Economics.

SHERIDAN, J. E. (1992), «Organizational Culture and Employee Retention», *Academy of Management Journal*, 35(5), 1036-1056.

SEGARRA, A. y CALLEJÓN, M. (2002), «New Firm's Survival and Market Turbulence: New Evidence from Spain», *Review of Industrial Organization*, 20, 1-14.

SINGER, B. y SPILERMAN, S. (1976a), «The Representation of Social Processes by Markov Models», *American Journal of Sociology*, vol 82, págs. 1-54.

— (1976b), «Some Methodological Issues in the Analysis of Longitudinal Surveys», *Annals of Economic and Social Measurement*, vol 5, págs. 447-474.

SINGER, Judith D. y WILLETT, John B. (1993), «It's About Time: Using Discrete-Time Survival Analysis to Study Duration and the Timing of Events», *Journal of Educational Statistics*, verano, 1993, 18, 155-195.

STATA CORP. (2001), *Stata Statistical Software: Release 7.0.* College Station, TX, Stata Corporation.

STRUTHERS, C. A. y KALBFLEISCH, J. D. (1984), «Misspecified Proportional Hazard Models», *Biometrica*, 73, 363-369.

SUÁREZ, F. F. y UTTERBACK, J. M. (1995), «Dominant designs and the survival of firms», *Strategic Management Journal*, vol. 16(6), págs. 415-430.

THERNEAU, T. M. (1997), «Extending the Cox Model», *Proceedings of the First Seattle Symposium in Biostatistics*, Nueva York, Springer-Verlag.

THERNEAU, T. M. y HAMILTON, S. A. (1997), «rhDNase as an example of recurrent event analysis», *Statistics in Medicine*, 16, 2029-2047.

THOMPSON, J. D. (1967), *Organizations in Action*, Nueva York, McGraw-Hill.

TREVOR, C. O. (2001), «Interactions Among Actual Ease of Movement Determinants and Job Satisfaction in the Prediciton of Voluntarily Turnover», *Academy of Management Journal*, 44(4), 621-638.

VERMUNT, J. K. (1997), *Log-linear models for event histories* (Advanced quantitative techniques in the social sciences, vol. 8), Thousand Oaks, Ca.: Sage.

WEI, L. J.; LIN, Y. y WEISSFELD (1989), «Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions», *Journal of the American Statistical Association*, 84 (408), 1065-1073.

YAMAGUCHI, K. (1991), *Event History analysis, Applied Social Research Method Series, 28,* Londres, Sage.

YOUNG, G.; SMITH, K. G.; GRIMM, C. M. y SIMON, D. (2000), «Multimarket contact and resource dissimilarity: A competitive dynamics perspective», *Journal of Management*, 26(6), 1217-1236.

ZORN, C. J. W. y VAN WINCKLE, S. R. (2000), «A Competing Risks Model of Supreme Court Vacancies, 1789-1992», *Political Behavior* 22(2), 145-166.