MODELO DE REGRESION DE COX: EJEMPLO NUMERICO DEL PROCESO DE ESTIMACION DE PARAMETROS

Alfonso Luis PALMER POL
Dpto. Psicología. Universidat de les Illes Balears

En los estudios longitudinales para el análisis del cambio, se recomienda utilizar el plan de observación denominado Event History Data. El modelo de riesgo proporcional permite estudiar el efecto de un conjunto de variables explicativas sobre la función de riesgo. En este contexto, se introduce el método de verosimilitud parcial que permite realizar la estimación de los parámetros del modelo de regresión de Cox. Se distinguen las tres situaciones posibles segun el tipo de datos: completos, incompletos y empates. Un ejemplo sencillo permite seguir, de forma numérica y exhaustiva, todos los pasos presentes en el proceso de estimación de parámetros, su significación y su interpretación.

Palabras Clave: Regresión de Cox; verosimilitud parcial; modelo de riesgo proporcional.

Cox regresion model: A numerical example of the parameters estimation process. In longitudinal studies for analyze change, we recomend to use the observation plan called Event History Data. The proportional hazard model allows to study the effect of a set of explicatives variables about the hazard function. In this context, we introduce the partial likelihood method that permits to estimate the parameters of the Cox regression model. We distinguish between three situation which are possible depending on the type of data: complete, incomplete and tie. A simple example will permit to follow in a numerical and exhaustive way, all the steps present at the parameters estimation process, their signification and interpretation.

Key Words: Cox regression; partial likelihood; proportional hazard model.

En los estudios longitudinales que tienen por objetivo el estudio del cambio se recomienda (Tuma y Hannan, 1984, p.22), para la recogida sistemática de la información pertinente, la utilización del plan de observación denominado Event-History Data. La ventaja de este diseño frente al clásico diseño de panel, es que permite obtener la máxima información de cada cambio producido, recogida por medio de la secuencia de cambio y del momento temporal en que éste se produce, y permite utilizar asimismo la información de las observaciones en las que no se ha producido ningún cambio (datos incompletos).

En esta situación, el *modelo de riesgo proporcional* es el modelo más utilizado para representar los efectos de un conjunto de variables explicativas sobre la variable tiempo de cambio (tiempo de supervivencia), o más bien sobre la probabilidad condicional de cambio, es decir sobre la función de riesgo $h(t)$. Suponemos que para cada sujeto tenemos un vector X de variables explicativas, concomitantes o pronóstico. Las componentes de dicho vector pueden representar tratamientos, definidos por medio de variables indicadoras (dummy variables), propiedades intrínsecas de los sujetos, tales como, por ejemplo, la edad, el sexo,

características individuales, agrupaciones cualitativas de los sujetos, o bien variables exógenas, como pueden ser las propiedades ambientales del problema.

El modelo de regresión de Cox (1972) viene determinado por la relación:

$$h(t/x) = h_0(t)\ \exp(X'B) \qquad (1)$$

donde la dependencia temporal está incluida en la tasa de riesgo de línea base, $h_o(t)$, y las variables concomitantes actúan en forma log-lineal, $exp(X'B)$, donde B es un vector de coeficientes de regresión desconocidos que parametrizan el modelo.

Este modelo puede describirse (Allison, 1984) como semiparamétrico o parcialmente paramétrico. Es paramétrico ya que especifica un modelo de regresión con una forma funcional específica; es no paramétrico en cuanto que no especifica la forma exacta de la distribución de los tiempos de supervivencia.

En este modelo las variables concomitantes actúan sobre la función de riesgo de forma multiplicativa. Las variables explicativas además pueden ser dependientes o independientes del tiempo.

El modelo de Cox puede utilizarse en los siguientes casos:

—Cuando no se tiene información previa acerca de la dirección temporal de la función de riesgo.

—Cuando siendo conocida la dirección, no puede ser determinada por un modelo paramétrico.

—Cuando se está únicamente interesado en la magnitud y la dirección de los efectos de las variables concomitantes, teniendo controlada la dirección temporal.

Debido a la existencia de datos incompletos, los parámetros del modelo de Cox no pueden ser estimados por el método ordinario de máxima verosimilitud al ser desconocida la forma específica de la función arbitraria de riesgo. El objetivo de este artículo es introducir el método de estimación del modelo de regresión de Cox y realizar una aplicación numérica del proceso de estimación de los parámetros del modelo de Cox, lo cual no se encuentra habitualmente en los textos y pensamos que es un ejercicio muy recomendable ya que permite tener una idea más clara del funcionamiento de dicho proceso.

Cox (1975) propuso un método de estimación denominado *verosimilitud parcial* (partial likelihood), siendo las verosimilitudes condicionales y marginales casos particulares del anterior.

El método de verosimilitud parcial se diferencia del método de verosimilitud ordinario en el sentido de que mientras el método ordinario se basa en el producto de las verosimilitudes para todos los individuos de la muestra, el método parcial se basa en el producto de las verosimilitudes de todos los *cambios* ocurridos.

Para estimar los coeficientes B en el modelo de Cox, en ausencia de conocimiento de $h_o(t)$, éste propuso la siguiente función de verosimilitud:

$$L(B) = \prod^{k} [\exp(XB) \ / \ \Sigma\exp(X\ B)] \qquad (2)$$

Esta expresión $L(B)$ no es una verdadera función de verosimilitud ya que no puede derivarse como la probabilidad de algún resultado observado bajo el modelo de estudio, si bien, como indica Cox (1975), puede tratarse como una función de verosimilitud ordinaria a efectos de realizar estimaciones de B.

Dichas estimaciones son consistentes (Cox, 1975; Tsiatis, 1981) y eficientes (Efron, 1977).

Cuando en un mismo instante t_i se produce más de un cambio, lo cual puede ocurrir cuando la variable tiempo se mide de forma discreta, la probabilidad de ocurrencia de los d_i cambios observados, condicionados al conjunto de riesgo R_i, viene dado (Cox, 1972) por:

$$\exp(Z'B) \ / \ \Sigma(Z'B)$$

donde cada elemento z_j del vector Z es la suma de los valores x_j sobre los d_i individuos que realizan un cambio en el instante ti y la suma del denominador se efectúa sobre los R_i sujetos expuestos al riesgo en t_i.

El logaritmo de la función de verosimilitud parcial viene dada, entonces, por:

$$LnL(B) = \Sigma(Z_i B) - \Sigma[d_i * Ln\ (\Sigma(\exp(B'X)))] \qquad (3)$$

Las estimaciones máximo verosímiles de *B* son estimaciones que maximizan la función *LnL(B)*.

Para estudiar el proceso de estimación de parámetros en el modelo de regresión de Cox vamos a utilizar un ejemplo simulado sobre un Programa de tratamiento libre de drogas, mediante un enfoque bio–psico–social.

El tratamiento de drogodependencias es un proceso continuo formado por diferentes fases y programas. Se empieza por un Programa de preparación al tratamiento y diagnóstico, seguido por un Programa de desintoxicación. A continuación se procede a un Programa de deshabituación, tras del cual se lleva a cabo el Programa de reinserción socio-laboral y finalmente se realiza un Programa de seguimiento.

El programa de desintoxicación, llevado a cabo desde el marco ambulatorio, puede realizarse con medicación o bien por medio de un soporte psico–social del entorno.

A continuación se presenta una matriz de datos que incluye las variables necesarias para nuestro análisis. La variable TIEMPO define, en semanas, el tiempo que un heroinómano ha estado incluido en el programa; la variable ESTADO define la situación del sujeto en su última observación (0=Sano, 1=Recaída), obtenida por determinación analítica de dro-

MATRIZ DE DATOS

CASO	TIEMPO	ESTADO	GRUPO	EDAD
1	3	1	1	45
2	5	0	1	20
3	5	0	1	39
4	8	1	1	51
5	8	1	0	30
6	13	0	0	35
7	16	1	1	44
8	19	0	1	28
9	20	1	0	35
10	20	0	0	35
11	21	1	0	38
12	25	0	0	24

gas en la orina; la variable GRUPO diferencia a los sujetos según el tipo de desintoxicación seguida (0=Psico: Con soporte psico-social, 1=Farma: Con medicación). La EDAD del sujeto está medida en años.

Para mayor comodidad, los sujetos se han ordenado en función de la variable tiempo de inclusión en el programa.

La figura 1 muestra el esquema gráfico de los datos del ejemplo (o=incompleto, x=cambio) que permite visualizar el momento en que se han producido los cambios, así como saber, en cada instante t_i, el conjunto de riesgo R_i.

El primer paso consiste en calcular la función de verosimilitud que viene dada en función de los parámetros desconocidos y de los datos observados y el segundo paso es hallar el máximo de dicha función, el cual vendrá determinado por un conjunto concreto de valores de los parámetros.

Utilizando los datos del ejemplo, podemos construir la Tabla 1.

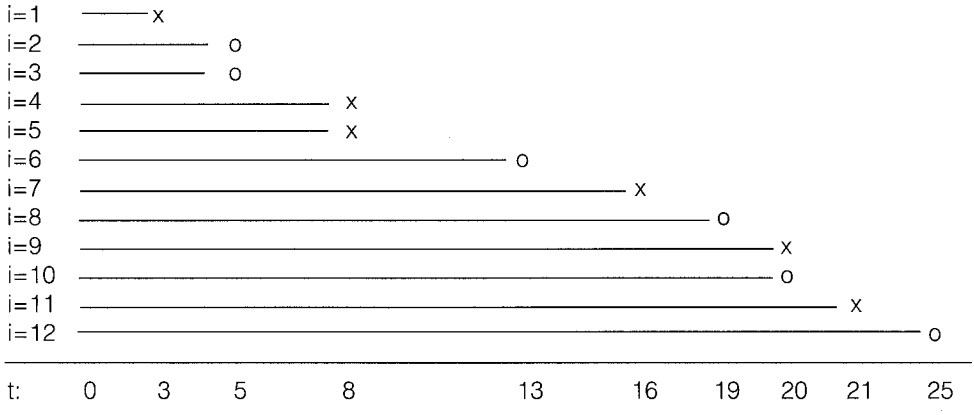

Figura 1. Esquema gráfico de los datos.

t_i	i	k	$R(t_i)$	Sujeto que cambia	Elementos del conjunto de riesgo
3	(1)	1	12	(1)	(1), (2), (3), (4), (5), (6), (7), (8), (9), (10), (11), (12)
5*	(2) (3)				
8	(4) (5)	3	9	(4)(5)	(4), (5), (6), (7), (8) (9), (10), (11), (12)
13*	(6)				
16	(7)	4	6	(7)	(7),(8),(9),(10), (11), (12)
19*	(8)				
20	(9)	5	4	(9)	(9),(10),(11),(12)
20*	(10)				
21	(11)	6	2	(11)	(11),(12)
25*	(12)				

Tabla 1. Expuestos al riesgo y cambios en los datos del ejemplo.

En la Tabla 1 se han ordenado los tiempos observados en la muestra en forma ascendente, señalando con un asterisco aquellos tiempos incompletos. Así pues, t_i representa el tiempo observado para el sujeto (i). En la tercera columna los valores de k representan los cambios ocurridos hasta el tiempo t_i inclusive. La cuarta columna representa el número de sujetos expuestos al riesgo de un cambio en cada tiempo t_i. La quinta columna identifica a los sujetos que han realizado un cambio en el instante t_i y la sexta columna enumera a los sujetos expuestos al riesgo en cada instante en el que ocurre un cambio.

La función de verosimilitud parcial (PL) viene dada por el producto de las siguientes probabilidades condicionales:

$$Pr[i=k/R(t_i)]$$

aplicadas a cada uno de los k cambios observados en la muestra estudiada. Es decir:

$$PL = Pr[i=1/R=12] * Pr[i=2/R=9] * Pr[i=3/R=8] * Pr[i=4/R=6] * Pr[i=5/R=4] * Pr[i=6/R=2] \quad (4)$$

La función de densidad en un instante t viene dada (Lee, 1980; Miller, 1981; Lawless, 1982), por:

$$f(t) = h(t)S(t) \quad (5)$$

donde *f(t)* representa la probabilidad de que ocurra el cambio en el instante *t* y *S(t)* representa la probabilidad de que el cambio se produzca pasado el tiempo t.

Cada uno de los términos de la expresión [4] que permite obtener la función de verosimilitud parcial, *PL*, puede expresarse en términos de funciones de riesgo.

Para comprobarlo procedemos a calcular uno de los elementos de la expresión [4]. Por ejemplo, el valor de Pr[i=5/R=4].

En el instante t=20, correspondiente a k=5, hay cuatro elementos expuestos al riesgo de un cambio: R(t=20) = (9), (10), (11), (12).

La probabilidad de que el cambio le ocurra a j=(9) en lugar de a j=(10) ó j=(11) ó j=(12) es:

$$f_9(20)\, S_{10}(20)\, S_{11}(20)\, S_{12}(20) = h_9(20)\, S_9(20)\, S_{10}(20)\, S_{11}(20)\, S_{12}(20)$$

Igualmente la probabilidad de que el cambio le ocurriera a j=(10) viene dada por:

$$S_9(20)\, f_{10}(20)\, S_{11}(20)\, S_{12}(20) = S_9(20)\, h_{10}(20)\, S_{10}(20)\, S_{11}(20)\, S_{12}(20)$$

y que le ocurriera al sujeto j=(11):

$$S_9(20)\, S_{10}(20)\, f_{11}(20)\, S_{12}(20) = S_9(20)\, S_{10}(20)\, h_{11}(20)\, S_{11}(20)\, S_{12}(20)$$

y que le ocurriera a j=(12):

$$S_9(20)\, S_{10}(20)\, S_{11}(20)\, f_{12}(20) = S_9(20)\, S_{10}(20)\, S_{11}(20)\, h_{12}(20)\, S_{12}(20)$$

Así pues:

$$\Pr[i=5/R=4] = \frac{f_9(20)S_{10}(20)S_{11}(20)S_{12}(20)}{f_9(20)S_{10}(20)S_{11}(20)S_{12}(20)+S_9(20)f_{10}(20)S_{11}(20)S_{12}(20)+S_9(20)S_{10}(20)f_{11}(20)S_{12}(20)+S_9(20)S_{10}(20)S_{11}(20)f_{12}(20)}$$

sustituyendo cada término por las igualdades anteriores se comprueba cómo se simplifica $S_9(20)S_{10}(20)S_{11}(20)S_{12}(20)$ del numerador y del denóminador quedando finalmente:

$$= \frac{h_9(20)}{h_9(20)+h_{10}(20)+h_{11}(20)} = \frac{\exp(X'B)}{\Sigma(\exp(X'B))}$$

una expresión que depende únicamente de los valores de las variables observadas, expresadas en términos de exponenciales.

Cada uno de los términos de la expresión [4] de *PL* viene determinado siguiendo este mismo proceso, de manera que la expresión general de la función de verosimilitud parcial será el producto de cada una de las expresiones halladas.

Podemos distinguir, siguiendo a Cox y Oakes (1984, p.91), tres situaciones distintas en la obtención de la función de verosimilitud parcial:

Caso de datos completos sin empates.

Este caso es realista en cuanto la variable tiempo de supervivencia tenga una distribución continua y su valor se registre de forma exacta, desapareciendo la posibilidad de empates.

Sean $t_1<t_2<...<t_n$ los *n* tiempos ordenados correspondientes a los n sujetos del estudio, sea $R(t_j)=[i:t_i \geq t_j]$ el conjunto de riesgo justo antes del tiempo t_j y sea r_j el tamaño de dicho conjunto. La figura 2 ilustra dichos conceptos.

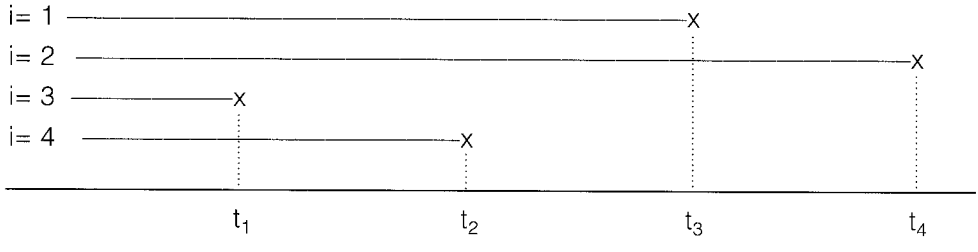

Figura 2. Esquema de datos completos sin empates.

En la figura 2 podemos ver los tiempos de supervivencia de 4 sujetos: $t_1< t_2< t_3< t_4$ siendo $R(t_1)$ = [1 2 3 4], $R(t_2)$ = [1 2 4], $R(t_3)$ = [1 2] y $R(t_4)$ = [2] los conjuntos de riesgo en cada tiempo.

La probabilidad p(3,4,1,2) de obtener la configuración conjunta de la figura 2 puede obtenerse por medio de la regla de la cadena para probabilidades condicionales:

$$p(3, 4, 1, 2) = p(3/1,2,3,4) * p(4/1,2,4) * p(1/1,2) * p(2/2)$$

Tal como hemos visto anteriormente, cada uno de estos elementos puede expresarse en términos de exponenciales. Así pues:

$$p(3/1,\ 2,\ 3,\ 4,) = \frac{Exp(3)}{Exp(1) + Exp(2) + Exp(3) + Exp(4)}$$

$$p(4/1,\ 2,\ 4) = \frac{Exp(4)}{Exp(1) + Exp(2) + Exp(4)}$$

$$p(1/1,\ 2) = \frac{Exp(1)}{Exp(1) + Exp(2)}$$

$$p(2/2) = \frac{Exp(2)}{Exp(2)}$$

De manera que:

$$p\ (3,\ 4,\ 1,\ 2) = \prod^{n} p(i/R(t_i)) = \prod^{n} \frac{Exp(i)}{\Sigma\ [Exp\ (R)]}$$

donde R representa a cada uno de los sujetos expuestos al riesgo en cada tiempo t_i.

Caso de datos incompletos

En este caso supongamos que tenemos d desenlaces (cambios, sucesos) observados en la muestra de tamaño n y ordenamos los tiempos $t_1 < t_2 < \ldots < t_d$. La figura 3 ilustra este caso:

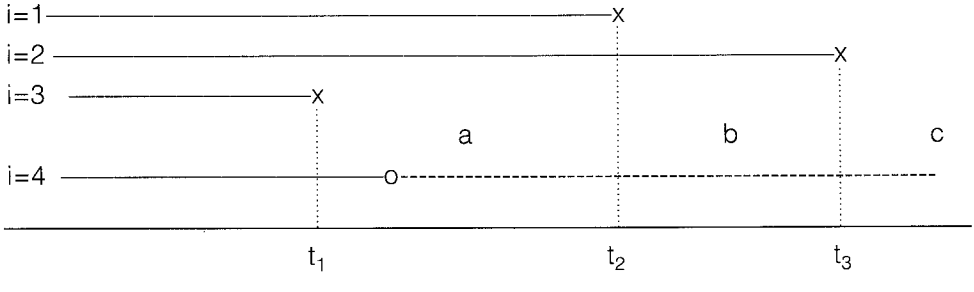

Figura 3. Caso de datos incompletos

En la figura 3 podemos ver los tiempos de 4 sujetos con un dato incompleto y tres datos completos: $t_1 < t_2 < t_3$ siendo $R(t_1) = [1\ 2\ 3\ 4]$, $R(t_2) = [1\ 2]$ y $R(t_3) = [2]$ los conjuntos de riesgo en cada tiempo, y a, b, c son las posibles posiciones que puede tomar el tiempo de la observación incompleta.

La probabilidad p(3,1,2) de obtener la combinación de desenlaces observada en la figura 3, siguiendo el esquema anterior, viene dada por el producto de probabilidades condicionadas al conjunto de riesgo que, como hemos visto anteriormente, puede expresarse en términos de exponenciales:

$$p(3,1,2) = \prod^{d} p(i/R(t_i)) = \prod^{d} \frac{Exp(i)}{\Sigma[Exp(R)]}$$

de manera que:

$$p(3,1,2) = p(3/1,2,3,4) * p(1/1,2) * p(2/2)$$

$$p(3,1,2) = \frac{Exp(3)}{Exp(1) + Exp(2) + Exp(3) + Exp(4)} \times \frac{Exp(1)}{Exp(1) + Exp(2)} \times \frac{Exp(2)}{Exp(2)}$$

Esta probabilidad puede obtenerse asimismo como suma de todas las probabilidades condicionales consistentes con el patrón de desenlaces y datos incompletos observados. Es decir:

$$p(3, 1, 2) = p(3, 4, 1, 2) + p(3, 1, 4, 2) + p(3, 1, 2,4)$$

El cálculo de cada sumando se realiza siguiendo el método descrito en el apartado anterior, a partir del cual se obtiene que:

$$P(3, 4, 1, 2) = \frac{Exp(3)}{Exp(1) + Exp(2) + Exp(3) + (Exp(4)} \times \frac{Exp(4)}{Exp(1) + Exp(2) + Exp(4)} \times$$

$$= \frac{Exp(1)}{Exp(1) + Exp(2)} \times \frac{Exp(2)}{Exp(2)}$$

$$p(3, 1, 4, 2) = \frac{Exp(3)}{Exp(1) + (Exp(2) + Exp(3) + (Exp(4)} \times \frac{Exp(1)}{Exp(1) + Exp(2) + Exp(3)} \times$$

$$\times \frac{Exp(4)}{Exp(2) + Exp(4)} \times \frac{Exp(2)}{Exp(2)}$$

$$p(3, 1, 2, 4) = \frac{Exp(3)}{Exp(1) + Exp(2) + Exp(3) + Exp(4)} \times \frac{Exp(1)}{Exp(1) + Exp(2) + (Exp(3)} \times$$

$$\times \frac{Exp(2)}{Exp(2) + Exp(4)} \times \frac{Exp(4)}{Exp(4)}$$

Caso de empates

La verosimilitud hallada en el apartado anterior, para datos incompletos, no es apropiada para distribuciones discretas del tiempo de supervivencia en las cuales podemos hallar empates en dichos tiempos.

El cálculo de la función de verosimilitud puede obtenerse sumando todos los términos de la función de verosimilitud marginal que son consistentes con los datos observados.

Así pues si las observaciones 1 y 2 realizan un cambio en el tiempo t siendo el conjunto de riesgo las observaciones 1, 2, 3 y 4, la contribución en la función de verosimilitud sería:

$$\frac{Exp(1)}{Exp(1) + Exp(2) + Exp(3) + Exp(4)} \times \frac{Exp(2)}{Exp(2) + Exp(3) + Exp(4)} +$$

$$+ \frac{Exp(2)}{Exp(1) + Exp(2) + Exp(3) + Exp(4)} \times \frac{Exp(1)}{Exp(1) + Exp(3) + Exp(4)}$$

Para simplificar los términos se aumentan todas las sumas de los denominadores para todas las R observaciones del conjunto de riesgo, lo cual permite escribir la expresión anterior como:

$$\frac{2 * Exp(1)\ Exp(2)}{[Exp(1) + Exp(2) + Exp(3) + Exp\ (4)]^2}$$

Esta expresión puede generalizarse para el caso de que existan d desenlaces en un instante *t*:

$$\frac{d! * Exp(1)... Exp(d)}{[\Sigma\ (Exp\ (R))]^d}$$

ESTIMACION DE LOS PARAMETROS

Para no complicar excesivamente los cálculos y las expresiones numéricas, y a efectos de una mayor simplicidad en la exposición, vamos a realizar el análisis utilizando únicamente una variable explicativa: la edad del sujeto en años (En Hill et al. (1990) puede encontrarse otras formas de codificar una variable cuantitativa).

Así pues, lo que se pretende es averiguar cómo afecta la edad del sujeto en el riesgo de que un heroinómano tenga una recaída en su tratamiento de desintoxicación. A continuación vamos a exponer los distintos pasos que se deben realizar para obtener la estimación de los parámetros del modelo.

Cálculo del logaritmo de la función de verosimilitud parcial

Sean $t_1 < t_2 < ... < t_k$ los *k* tiempos diferentes en los que ocurre algún cambio entre los *n* tiempos de supervivencia.

Las estimaciones máximo-verosímiles se obtienen maximizando la función:

$$LnL(B) = \sum^{k} (Z_i\, b) - \sum^{k} [d_i * Ln\, (\sum_{R_i} (exp\, (X'b)))]$$

siendo Z_i la suma de los valores de la variable concomitante para todos los d_i individuos que realizan un cambio en el instante t_i.

Desarrollando la expresión general para cada uno de los 5 instantes en los que se produce algún cambio, se obtiene el Logaritmo de la función de verosimilitud, en función del parámetro que se desea estimar:

$$\begin{aligned}LnL(B) = \quad & 45b - 1*Ln\,[\exp(45b) + \exp(20b) + ... + \exp(38b) + \exp(24b)] \\ & + 81b - 2*Ln\,[\exp(51b) + \exp(30b) + ... + \exp(38b) + \exp(24b)] \\ & + 44b - 1*Ln\,[\exp(44b) + \exp(28b) + ... + \exp(38b) + \exp(24b)] \\ & + 35b - 1*Ln\,[\exp(35b) + \exp(35b) + \exp(38b) + \exp(24b)] \\ & + 38b - 1*Ln\,[\exp(38b) + \exp(24b)]\end{aligned}$$

Cálculo del vector de primeras derivadas

Las derivadas parciales del logaritmo de la función de verosimilitud parcial, respecto a cada uno de los coeficientes, se calculan de acuerdo a las siguientes ecuaciones:

$$U_k(B) = \sum^{k} [Z_i - m_i * \sum_{R_i} (X\exp(X'b)) / \sum_{R_i} (\exp(X'b))]$$

El sumatorio general se realiza para los *k* instantes en los que se produce algún desenlace. Los otros sumatorios se realizan para los sujetos expuestos al riesgo.

Desarrollando la expresión general para cada uno de los 5 instantes en los que se produce algún cambio, se obtiene el vector de primeras derivadas, en función del parámetro que se desea estimar:

$$U(B) = 45 - \frac{45\exp(45b) + ... + 24\exp(24b)}{\exp(45b) + \exp(20b) + ... + \exp(24b)} +$$

$$81 - 2 * \frac{51\exp(51b) + ... + 24\exp(24b)}{\exp(51b) + \exp(30b) + ... + \exp(24b)} +$$

$$44 - \frac{44\exp(44b) + ... + 24\exp(24b)}{\exp(44b) + \exp(28b) + ... + \exp(24b)} +$$

$$35 - \frac{35\exp(35b) + ... + 24\exp(24b)}{\exp(35b) + \exp(35b) + ... + \exp(24b)} +$$

$$38 - \frac{38\exp(38b) + 24\exp(24b)}{\exp(38b) + \exp(24b)}$$

Cálculo de la matriz de segundas derivadas

Las segundas derivadas parciales del logaritmo de la función de verosimilitud parcial, respecto a cada uno de los coeficientes, se calculan de acuerdo a las siguientes ecuaciones:

$$I_{kk'}(b) = \sum^{k} [m_i * [\sum_{R_i} (X_k X_{k'} \exp(X'b)) / \sum_{R_i} (\exp(X'b)) - (\sum_{R_i} (X_k \exp(X'b)))$$

$$(\sum_{R_i} (X_{k'} \exp(X'b))) / (\sum_{R_i} (\exp(X'b))^2))]]$$

Desarrollando la expresión general para cada uno de los 5 instantes en los que se produce algún cambio, se obtiene la matriz de segundas derivadas, en función del parámetro que se desea estimar:

$$I(B) = \left[\frac{45^* 45\exp(45b) + \ldots + 24^* 24\exp(24b)}{\exp(45b) + \exp(20b) + \ldots + \exp(24b)} - \frac{[45\exp(45b) + \ldots + 24\exp(24b)]^*[45\exp(45b) + \ldots + 24\exp(24b)]}{([\exp(45b)+\exp(20b)+\ldots+\exp(24b)^2} \right] + \ldots +$$

$$\left[\frac{38^* 38\exp(38b) + 24^*24\exp(24b)]}{\exp(38b) + \exp(24b)} - \frac{[38\exp(38b) +24\exp(24b)] * [38\exp(38b) + 24\exp(24b)]}{([\exp(38b) + \exp(24b)]^2)} \right]$$

ESTIMACION DEL COEFICIENTE B (METODO DE NEWTON-RAPHSON): CALCULO DEL VALOR B* EN CADA ITERACION

1) Se realiza una estimación inicial $b=b_0$. Generalmente se toma cero como primer valor. Así pues, la primera estimación es b=0.

2) Se calculan los valores de $U(b_0)$ y de $I(b_0)$: Para ello se sustituye el valor b=0 en las expresiones anteriormente halladas del vector de primeras derivadas U(B) y de la matriz de segundas derivadas I(B).

Cálculo de U(0)

Sustituyendo el valor inicial b=0 en la expresión general del vector de primeras derivadas U(b), se obtiene el valor:

$$U(0) = (45 - 424/12) + (81 - 2*320/9) + (44 - 204/6) + (35 - 132/4) + (38 - 62/2) = 38.5555$$

Cálculo de I(0)

Sustituyendo el valor inicial b = 0 en la expresión general de la matriz de segundas derivadas I(b), se obtiene el valor:

$$\begin{aligned} I(0) = \; & (15862/12 - (424*424)/(12^2)) \\ & +2(11916/9- (320*320)/(9^2)) \\ & + (7190/6- (204*204)/(6^2)) \\ & + (4470/4- (132*132)/(4^2)) \\ & + (2020/2-(62*62)/(2^2)) \quad = 312.827 \end{aligned}$$

3) Se calcula la siguiente aproximación b* de B, por medio de la expresión:

$$b^* = b_0 + I^{-1}(b_0)U(b_0)$$

La segunda estimación viene dada por:

$$b = 0 + (38.555/312.827) = 0.12324881$$

4) Se repiten los pasos 2 y 3 reemplazando b_0 por b*.

$$U(b) = 45 - \frac{68394.50347}{1577.160825} + 81 - 2 * \frac{51856.71381}{1186.808914} + 44 - \frac{20652.72618}{534.9127812} +$$

$$+35 - \frac{9801.841895}{276.8360083} + 38 - \frac{4571.648199}{127.4019026} = 1.634414 - 6.388480 + 5.390479 - 0.406673 +$$

$$+ 2.116327 = 2.346067175$$

$$I(b) = \frac{3047460.317}{1577.160825} - \frac{(68394.5) * (68394.5)}{(1577.16)^2}$$

$$+2 \left[\frac{2337766.067}{1186.808914} - \frac{(51856.7) * (51856.7)}{(1186.8)^2} \right]$$

$$+ \frac{813621.5166(20662.7) * (2066.27)}{534.9127812} - \frac{}{(534.9)^2}$$

$$+ \frac{350308.4325}{276.8360083} - \frac{(9801.8) * (9801.8)}{(276.8)^2}$$

$$+ \frac{167251.6532}{127.4019026} \frac{(4571.6) * (4571.6)}{(127.4)^2}$$

$$= 51.6704386 + 2* 60.6048327 + 30.3407495 + 11.7680724 + 25.1497456 = 240.1386715$$

La siguiente iteración proporciona la estimación:

$$b = 0.12324881 + \frac{2.346067175}{240.1386715} = 0.133018445$$

A partir de este valor, y siguiendo los pasos anteriores, se llega, en la última iteración al valor b=1.33293.

5) En cada iteración j se estudia la convergencia, finalizando el proceso cuando:

$$\frac{LnL(B_j) - LnL(B_{j-i})}{LnL(B_j)} < c$$

siendo c el criterio de convergencia (el BMDP utiliza el valor c=0.00001).

CALCULOS FINALES PARA EL COEFICIENTE OBTENIDO

Una vez obtenido el valor final b = 1.33293, se puede estimar la variabilidad de la distribución muestral del coeficiente, así como proceder al estudio de la significación global del modelo.

Valor del logaritmo de la función de verosimilitud parcial

$$
\begin{aligned}
\text{LnL(B)=} \quad & 5.9985 - \text{Ln}[402.82 + 14.38 + 181.03 + 158.44+24.51] \\
& + 10.7973 - 2*\text{Ln}[896.32 + 54.54 + 106.2 + 1\,58.44+24.51] \\
& +5.8652 - \text{Ln}[352.55 + 41.78 + 158.44 + 24.51] \\
& +4.6655 - \text{Ln}[106.2 + 106.2 + 158.44 + 24.51] \\
& +5.0654 - \text{Ln}[158.44 + 24.51] = \\
& \\
& = 5.9985 - \text{Ln}[2445] \\
& + 10.7973 - 2*\text{Ln}[1845.71] \\
& + 5.8652 - \text{Ln}[788.65] \\
& + 4.6655 - \text{Ln}[394.32] \\
& + 5.0654 - \text{Ln}[181.92] = -8.3132
\end{aligned}
$$

Cálculo del error estandar del coeficiente B hallado

Puesto que el valor hallado del coeficiente *b* es una estimación del valor real, vendrá afectado por un determinado error. El valor del error estándar puede ser obtenido por medio de la matriz de información de Fisher. Para ello se obtiene el valor de la matriz de segundas derivadas para el coeficiente b = 0.1333 obtenido.

Este valor resulta ser: I(b) = 226.81

La matriz de variancias-covariancias de los coeficientes viene dada por la inversa de la matriz de información. En nuestro caso, al ser un único coeficiente, la variancia viene dada por la inversa del valor obtenido:

$$I(B)^{-1} = \text{Var}(b) = 1/226.81 = 0.004409$$

De donde resulta que el error estándar buscado vale:

$$SE(b) = 0.0664$$

Cálculo del test JI-Cuadrado global

A partir del vector de primeras derivadas y de la matriz de segundas derivadas, el test de Rao (*score test*) permite estudiar la hipótesis nula de que el vector de coeficientes es un vector nulo. Bajo la hipótesis nula este test (Miller, 1981, p.125; Lawless, 1982, p.440; Hill et al., 1990, p.83) sigue una distribución Ji-cuadrado con grados de libertad igual al orden del vector de coeficientes estimado.

En nuestro caso, el orden del vector es 1. Así pues:

$$X^2 = U'(0)\, I^{-1}(0)\, U(0) = 38.55 * \frac{1}{312.827} * 38.55 = 4.75$$

Valor significativo ($P<0.05$) al ser inferior al valor límite 3.84.

INTERPRETACION DEL COEFICIENTE DEL MODELO DE REGRESION DE COX

Los coeficientes del modelo de regresión de Cox indican la relación existente entre la variable explicativa correspondiente y la función de riesgo. Un valor positivo del coeficiente supone un aumento en el valor de la función de riesgo para el sujeto, lo cual conlleva una relación negativa con el tiempo de cambio. Es decir, un coeficiente positivo indica un mayor riesgo de que se produzca el cambio.

La interpretación de b es paralela a la de un coeficiente de regresión no estandarizado. El modelo de regresión de Cox de nuestro ejemplo viene dado por:

$$h\,(t/x\,) = h_o(t)\,\exp\,(0.1333^*\,EDAD)$$

El valor *b=0.1333* significa que el incremento de un año de edad aumenta el logaritmo de la tasa de riesgo en 0.1333 (controlando las otras variables incluidas en la ecuación).

Una interpretación más intuitiva se obtiene a partir del valor $e^b = 1.1426$: Para cada unidad de incremento en la variable explicativa EDAD, la tasa de riesgo se multiplica por 1.1426.

Además, $100\,(e^b-1) = 100(1.1426-1) = 14.26$ proporciona el cambio en porcentaje en la tasa de riesgo que se produce con cada unidad de cambio en la variable independiente. Es decir, que cada año aumenta la tasa de riesgo un 14.26% respecto a su valor en el año anterior.

REFERENCIAS

Allison, P.D. (1984). *Event history analysis.* Sage University Paper, 07-046. London: Sage Publications.

Blossfeld, H.P., Hamerle, A. y Mayer, K.U. (1989). *Event history analysis.* Hillsdale, NJ: LEA.

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

Carter, W.H., Wampler, G.L. y Stablein, D.M. (1983). *Regression analysis of survival data in cancer chemotherapy.* N.Y.: Marcel Dekker.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, 34, 187-202.

Cox, D.R. (1975). Partial likelihood. *Biometrika*, 62, 269-276.

Cox, D.R. y Oakes, D. (1984). *Analysis of survival data.* London: Chapman and Hall.

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

Elandt-Johnson, R.C. y Johnson, N.L. (1980). *Survival models and data analysis.* New York: John Wiley and Sons.

Hill, C., Com-Nougué, C., Kramar, A., Moreau, T., O'quigley, J., Senoussi, R. y Chastang C. (1990). *Analyse statistique des données de survie.* Paris: Flammarion.

Kalbfleisch, J.D. y Prentice, R.L. (1980). *The statistical analysis of failure time data.* N.Y.: John Wiley and Sons.

Lawless, J.F. (1982). *Statistical models and methods for lifetime data.* N.Y.: John Wiley and Sons.

Lee, E.T. (1980). *Statistical methods for survival data analysis.* Calif: Lifetime Learning Publications.

Leurgans, S. (1980). Exploring the influence of several factors on a set of censored data. En R.G. Miller, B. Efron, B.W. Brown, L.E. Moses (eds.), *Biostatistics casebook.* N.Y.: John Wiley and Sons.

Miller, R.G. (1981). *Survival analysis.* N.Y.: John Wiley and Sons.

Nelson, W. (1982). Applied life data analysis. N.Y.: John Wiley and Sons.

Palmer, A. (1985). *Análisis de la supervivencia.* Barcelona: Universidad Autónoma de Barcelona.

Palmer, A. y Losilla, J.M. (1991). El modelo de azar proporcional: la regresión de Cox. *Qurriculum*, extra 1/2, 57-61.

Tsiatis, A. (1981). A large sample study of Cox's regression model. *Annals of Statistics* 9, 93-108.

Tuma, N.B. y Hannan, M.T. (1984). *Social dynamics: Models and methods.* London: Academic Press.

ANEXO

LISTADO DEL MODELO DE REGRESION DE COX PARA EL ESTUDIO DE LA INFLUENCIA DE LA EDAD

A continuación se muestran los resultados del análisis realizado por medio del programa 2L del BMDP. La siguiente tabla reproduce los datos del ejemplo que hemos utilizado.

CASE NO	1 TIEMPO	2 ESTADO	3 GRUPO	4 EDAD
1	3	RECAIDA	FARMA	45
2	5	SANO	FARMA	20
3	5	SANO	FARMA	39
4	8	RECAIDA	FARMA	51
5	8	RECAIDA	PSICO	30
6	13	SANO	PSICO	35
7	16	RECAIDA	FARMA	44
8	19	SANO	FARMA	28
9	20	RECAIDA	PSICO	35
10	20	SANO	PSICO	35
11	21	RECAIDA	PSICO	38
12	25	SANO	PSICO	24

NUMBER OF CASES READ 12

A continuación se proporcionan algunos índices estadísticos sobre la variable explicativa que se va a introducir en el modelo de regresión de Cox.

DESCRIPTIVE STATISTICS FOR FIXED COVARIATES

VARIABLE NO. NAME	MINIMUM	MAXIMUM	MEAN	STANDARD DEVIATION	SKEWNESS	KURTOSIS
4 EDAD	20.0000	51.0000	35.3333	8.9477	-0.02	1.95

La siguiente tabla proporciona, sobre el total de sujetos, el número de observaciones en cada una de las categorías de la variable Estado.

STATUS CODE FREQUENCIES

TOTAL	RECAIDA	SANO	PERCENT SANO
12	6	6	0.5000

A continuación se muestran los valores del logaritmo de la función de verosimilitud parcial, así como el valor estimado del parámetro b, coeficiente de la variable independiente EDAD, en las diferentes iteraciones realizadas.

ITERATION NUMBER	HALVINGS	LOG LIKELIHOOD	PARAMETER ESTIMATES
0	0	-10.75055682	0.000000
1	0	-8.32484337	0.123249
2	0	-8.31318030	0.133018
3	0	-8.31317172	0.133293

Para el último parámetro estimado, el cual cumple el criterio de convergencia, se proporciona el valor del logaritmo de la función de verosimilitud parcial para el valor estimado del coeficiente, así como el valor de la prueba Ji-Cuadrado que permite estudiar la hipótesis nula según la cual el coeficiente vale cero, así como la significación del valor obtenido.

LOG LIKELIHOOD = -8.3132
GLOBAL CHI-SQUARE = 4.75 D.F.= 1 P-VALUE = 0.0293

La siguiente tabla proporciona los datos necesarios para estudiar e interpretar el coeficiente obtenido para la variable independiente utilizada.

VARIABLE	COEFFICIENT	STANDARD ERROR	COEFF./S.E.	EXP(COEFF.)
EDAD	0.1333	0.0664	2.0080	1.1426