Legalización y disolución de uniones consensuales: un ejemplo del uso de modelos log-lineales para estimar modelos de riesgos en competencia

José Gómez de León*

El trabajo que presentamos aquí es principalmente metodológico y consiste en hacer una generalización de los llamados modelos de riesgos proporcionales a los casos en donde intervienen múltiples riesgos y cuando éstos operan en competencia. Como veremos, los modelos de riesgos proporcionales son una extensión de la metodología implícita en el cálculo de una tabla de mortalidad, en donde la función de riesgo de la tabla se hace depender de algunas variables (covariables) a modo de modelo de regresión. Huelga decir que la metodología de la tabla de vida ha sido desde su origen en el siglo XVII (Halley, 1693) un instrumento esencial del cálculo y del análisis demográfico. El énfasis de este trabajo consiste en mostrar la utilidad de estos adelantos metodológicos, producidos más bien en el terreno de la bioestadística y de la ingeniería de sistemas, en el análisis de datos demográficos.

Introducción

La mayor parte de los eventos de que se ocupa la demografía pueden ser vistos –y de hecho, es una dimensión que siempre subyace en el análisis– como riesgos o probabilidades en competencia. El caso que a modo ilustrativo hemos decidido analizar aquí, la legalización de uniones consensuales, es un buen ejemplo de esto: una unión consensual puede terminar como tal, primero, por alguna forma de legalización o, segundo, por alguna forma de separación. Entre las legalizaciones podemos considerar las diversas combinaciones de formas de matrimonio, y entre las separaciones debemos considerar la viudez o la disolución de la unión. Puesto que todas estas opciones son excluyentes se dice que están "en competencia". Así, para estimar las probabilidades de legalización (libres del efecto perturbador que introduce, por ejemplo, la viudez) es necesario considerar el fenómeno como expuesto a dos procesos de decremento. Otros fenómenos demográficos, como la mortalidad vista por causas, suponen la acción de múltiples procesos (independientes) de decremento.

* Coordinador nacional del Programa de Educación, Salud y Alimentación (Progresa).

Al análisis de riesgos en competencia, los modelos de riesgos proporcionales añaden una dimensión, la de la dependencia de los riesgos de otras variables que en principio "explican" las diferencias en los niveles de riesgo. Antes de esta extensión (la de los modelos de riesgos proporcionales) un procedimiento común para estudiar diferenciales de riesgos, consistía en calcular tablas de mortalidad para cada combinación de los factores de riesgo y luego comparar las tablas entre sí. Este procedimiento, para producir resultados significativos, requiere de poblaciones muy numerosas y es factible sólo para un número muy limitado de factores de riesgo. Los modelos de riesgos proporcionales están orientados justamente a superar estas limitaciones y otras igualmente restrictivas, como es el análisis de datos truncados (*censored data*).

El trabajo está organizado como sigue. En la primera parte se describe "La naturaleza de los modelos de sobrevivencia" (o modelos de mortalidad, si se quiere), se dan algunas definiciones, y se establece la nomenclatura.

En la segunda y tercera secciones ("Modelos de regresión" y "Modelos log-lineales para datos categoriales") se describe, respectivamente, la forma general de los modelos de regresión para el análisis de riesgos simples, y su representación y manipulación como modelos log-lineales. En el siguiente apartado se introducen los "Modelos de riesgos en competencia con covariables". Posteriormente, se dan algunos elementos para la estimación y la interpretación de éstos ("Estimación y prueba de hipótesis"). Por último, se hace un análisis, mediante los métodos propuestos, de los determinantes de la "Legalización de uniones consensuales en México".

La naturaleza de los modelos de sobrevivencia

Notación y definiciones de base

La información básica para construir o estimar cualquier modelo de sobrevivencia consiste en disponer de una serie de observaciones sobre el *tiempo* (tiempo-duración) que tarda en ocurrir un evento dado, como puede ser la muerte, cambiar de residencia, la disolución de una unión, etc. Ese *tiempo* es en realidad el tiempo de *sobrevivencia* al evento en cuestión dada una condición particular, la de no haber sufrido el evento y estar expuesto al riesgo de sufrirlo.

En el caso típico, se requieren los datos de sobrevivencia de n individuos (independientes) observados durante un cierto lapso de tiempo. Consideremos ahora una variable aleatoria continua T que representa los tiempos de sobrevivencia de esos individuos, que de ahora en adelante vamos a considerar como si hubiesen sido seleccionados aleatoriamente de una población homógenea.

La función de sobrevivencia para cualquier momento t se define como la probabilidad de que T sea al menos tan grande como t, es decir,

$$S(t)=Pr(T\geq t), t\geq 0 \qquad [1]$$

que indica la probabilidad de que un individuo sobreviva hasta el momento t.

La probabilidad de que un individuo muera (no sobreviva al evento en cuestión) en el intervalo de tiempo $(t, t+\Delta t)$ (no importa cuan pequeño sea Δt) se define como:

$$f(t)=\lim_{\Delta t\to 0}\frac{Pr(t\leq T<t+\Delta t)}{\Delta t} \qquad [2]$$

que es la función de densidad probabilistica (f.d.p.) de la variable T.

Por último, la probabilidad condicional de que un individuo muera en el intervalo $(t, t+\Delta t)$, dado que ya sobrevivió hasta el momento t, se define como:

$$h(t)=\lim_{\Delta t\to 0}\frac{Pr(t\leq T<t+\Delta t|T>t)}{\Delta t}$$

$$=\frac{f(t)}{S(t)} \qquad [3]$$

que es la función de riesgo, también llamada tasa instantánea de mortalidad.

Las tres funciones $f(t)$, $S(t)$, y $h(t)$ caracterizan de modo equivalente a T, en términos de eventos, de sobrevivientes, y de probabilidades condicionales. Algunas relaciones matemáticas entre ellas son como sigue:

$$h(t) = -\frac{d \log S(t)}{dt} \qquad [4]$$

$$S(t) = \exp\left(-\int_0^t h(x)\,dx\right), \text{con } S(o) = 1 \qquad [5]$$

$$f(t) = h(t) \exp -\int_0^t h(x)\,dx \qquad [6]$$

Las expresiones [1] a [6] se cumplen también para el caso en que T sea una variable aleatoria discreta, y que tome sólo los valores $t_1, t_2, \ldots$ (con $0 < t_1, t_2 < \ldots$). En este caso, la función de riesgo se define como:

$$h(t_k) = Pr(T = t_k \mid T \geq t_k)$$

$$= \frac{f(t_k)}{S(t_k)}, \qquad k = 1, 2, \ldots$$

$$= 1 - (S(t_{k+1}) / S(t_k))$$

En el resto de este trabajo, la mortalidad va a ser caracterizada por la función de riesgo h(t) en el sentido de que las expresiones [5] y [6] nos permiten derivar las series de sobrevivientes y de decesos.

Modelos de mortalidad

El modelo paramétrico más sencillo de mortalidad es el h(t) = h, que especifica una tasa de mortalidad constante. A este modelo se le llama *exponencial* pues implica que f (t) = h exp (− h t) y S (t) = exp (− ht). Pese a su simplismo, mediante algunas adaptaciones, este modelo es extremadamente útil en demografía; de hecho, en la construcción convencional de una tabla de mortalidad se asume que $h(t) = h_k$ donde k son los intervalos de una función exponencial escalonada de sobrevivencia (Chiang, 1984).

Otros modelos paramétricos de mortalidad se basan en modelar alguna de las funciones [4] - [6] mediante distribuciones probabilísticas típicas, como la Weibull, la Log-normal, la Gamma, la Logística, etc. Para el caso que nos ocupa no es necesario revisar ninguna de es-

tas distribuciones; el lector interesado puede consultar Kalbfleisch y Prentice (1980).

Nuestro interés se restringe también a modelos paramétricos pues queremos precisamente caracterizar (mediante diferentes valores en los parámetros) la sobrevivencia de poblaciones heterogéneas. Los enfoques no paramétricos más comunes, como la estimación "producto-límite" de la función de sobrevivencia (Kaplan y Meier, 1958) suponen una población homogénea.

Modelos de regresión

Los modelos de regresión son candidatos naturales para estimar los "efectos" que tienen sobre la mortalidad diversos factores de riesgo. En efecto, supongamos que para cada individuo j, además del tiempo de sobrevivencia, se tiene también información sobre p covariables $z_j = (z_{1j}, \ldots, z_{pj})$. En este caso, un modelo en donde los riesgos $h(t)$ dependan de las covariables $\underline{z}$ puede formularse como (Glasser, 1967):

$$h(t,\underline{z}) = h\exp(\underline{z} \quad \underline{B}) \qquad [7]$$

que no es más que un análogo del modelo exponencial antes descrito, donde $\underline{B}$ es un vector de coeficientes de regresión.

En palabras, el modelo [7] especifica que el logaritmo de la función de riesgo es una función lineal de las covariables Z, es decir, el modelo [7] es un modelo log-lineal.

Un modelo más general que [7] (del que [7] es un caso particular) es el propuesto por Cox (1972):

$$h(t,\underline{z}) = h_o(t)\exp(\underline{z} \quad \underline{B}) \qquad [8]$$

donde $h_o(t)$ es una función de base (arbitraria) caracterizada cuando $\underline{z} = 0$ o $\underline{B} = 0$.

Gran parte de la flexibilidad (y la utilidad) del modelo [8] reside en cómo especificar la función $h_o(t)$. Más adelante tocaremos de nuevo este punto. Por lo pronto, podemos notar que la función de riesgo $h(t)$ es proporcional a la función de riesgo de base $h_o(t)$, de ahí que al modelo [8] se le llame de *riesgos proporcionales*. Así, la razón de

las funciones de riesgo de dos individuos con covariables $\underline{z}_1$ y $\underline{z}_2$ es independiente del tiempo-duración t y por lo tanto es una constante. En la notación de [8] tenemos que:

$$\frac{h(t,\underline{z}_1)}{h(t,\underline{z}_2)} = \exp[\ \underline{B}(\underline{z}_1 - \underline{z}_2)\ ] = \psi \qquad [9]$$

donde ψ es el riesgo relativo de morir para un individuo con covariables $\underline{z}_1$ comparado con un individuo con covariables $\underline{z}_2$ (Menken *et al.*, 1981; Schlesselman, 1982). Es obvio en la expresión [8] que el riesgo relativo ψ depende sólo de los factores de riesgo en los que difieren los dos individuos.

En principio, la función ho (t) puede ser modelada por cualquier distribución paramétrica que convenga. Sin embargo, para emparentar el modelo [8] con el modelo clásico de la tabla de mortalidad (al menos, tal como se les usa comúnmente en demografía) es conveniente definir a ho (t) como una función escalonada de la forma:

$$ho\ (t) = h_k ; t_k \leq t < t_{k+w_k}$$

donde k = 1, 2,... , K se refiere a los intervalos $I_k = (t_k, t_{k+w_k})$ (cada uno de duración w_k) que particionan a t. El modelo de sobrevivencia que resulta es entonces:

$$S(t_k|\underline{z}) = So(t_k)^{\exp(\underline{z}\ \underline{B})} \qquad [10]$$

donde S (t_k | $\underline{z}$) explicita que la función de sobrevivencia es escalonada y depende de $\underline{z}$, y donde So (t_k) es la función de sobrevivencia de base.

Las dos incógnitas del modelo [8] (o del modelo [10]) que requieren ser estimada son: los parámetros de regresión $\underline{B}$, y la función de riesgo ho (t). Un enfoque para estimar estas incógnitas es el propuesto por Cox (1975). En lugar de maximizar la función de verosimilitud de [8] simultáneamente con respecto a $\underline{B}$ y a ho (t) (como se haría en un problema convencional de estimación mediante máximo de verosimilitud), Cox propone estimar primero $\underline{B}$ (maximizando una función de verosimilitud "parcial", que no depende de ho (t)) para después estimar ho (t) a partir de $\underline{B}$. Bajo este enfoque, que es

el que seguimos aquí y se detalla un poco más adelante, la estimación de los parámetros $\underline{B}$ trae como subproducto la estimación de la función de base ho (t), es decir, de la tabla de mortalidad sin distinción de factores de riesgo.

Un último punto antes de dejar esta sección es el siguiente. Es obvio en [8] que las covariables $\underline{z}$ no dependen de la duración t; por ello a esta formulación se le denomina de riesgos proporcionales. En caso de que el efecto de algún(os) factor(es) de riesgo dependa de t, el modelo [8] puede ampliarse como sigue:

$$h[t,\underline{z}(t)]=ho(t)\exp[\underline{z}(t)\underline{B}] \qquad [11]$$

con lo cual [11] deja de ser un modelo de riesgos proporcionales. Cox (1975) y Kalbfleisch y Prentice (1980) muestran que el enfoque de "verosimilitud parcial" propuesto por Cox (1972) para estimar modelos de riesgos proporcionales sigue siendo válido para estimar casos como el del modelo [11], lo cual permite probar directamente sobre el modelo [8] (es decir, sin requerir procedimientos de estimación diferentes) si algunos factores de riesgo satisfacen la hipótesis de proporcionalidad o no.

Modelos log-lineales para datos categoriales

Bajo el supuesto de que las covariables estén medidas categóricamente, el modelo [9] puede rescribirse como:

$$\log h(t_k,\underline{z})=\log h_k+(\underline{z}\quad\underline{B});\ t_k\leq t<t_{k+w_k} \qquad [12]$$

que define, como dijimos, un modelo log-lineal (en las $\underline{z}$).

El modelo [12] puede a su vez rescribirse, utilizando la notación usual para modelos log-lineales, como Bishop *et al.* (1975) y Laird y Olivier, (1981):

$$\begin{aligned}\log h(t_{i_o},\underline{z}) &= \log\Theta_{i_o i_1\ldots i_p}\\ &= U+U_{0(i_o)}+U_{i(i_1)}+\ldots+U_{p(i_p)}+\\ &\quad +U_{01(i_o i_1)}+\ldots\end{aligned} \qquad [13]$$

$$+U_{01\ldots p}(i_o i_1 \ldots i_p)$$

donde i_o se refiere a los K intervalos de tiempo, e i_j (con j = 1,... , p) se refiere a las categorías de cada una de la p covariables. El modelo [13] está sujeto a las restricciones ANOVA habituales.

En el resto de esta sección, para simplificar la notación, se hace caso omiso de los índices i en el entendido de que se conoce el número de categorías de cada variable (incluida la participación de t). También para simplificar vamos a considerar un caso en el que la "mortalidad" se ve afectada sólo por dos factores de riesgo: 1 y 2.

La notación de [13] es particularmente útil para formular modelos de diversa complejidad. En el caso de nuestro ejemplo, algunos modelos de ln Θ (cuya pertinencia habría que probar estadísticamente) pueden ser los que siguen:

[I] U
[II] $U+U_0$
[III] $U+U_0+U_1$
[IV] $U+U_0+U_1+U_2$
[V] $U+U_0+U_1+U_2+U_{12}$
[VI] $U+U_0+U_1+U_2+U_{01}+U_{02}$
[VII] $U+U_0+U_1+U_2+U_{01}+U_{02}+U_{12}+U_{012}$

donde U es simplemente una constante; U_0 se refiere a los valores que toma la función-base de riesgo (ho (t) en continuo) en los K intervalos de tiempo; U_1 y U_2 se refieren a los efectos de las covariables 1 y 2 respectivamente; y U_{12} se refiere al efecto "conjunto" que ejercen las variables 1 y 2 sobre el logaritmo de $\Theta_{i_0 i_1 i_2}$.

El modelo *[I]* sería simplemente un modelo de sobrevivencia exponencial en el cual no hay covariables; si esta condición no se satisface y se requiere una función de riesgo escalonada, el modelo se convierte entonces en el modelo *[II]*. El resto de los modelos especifica diferentes combinaciones de factores de riesgo. Debemos notar que los modelos *[III]* a *[V]* especifican riesgos proporcionales, mientras que los modelos *[VI]* y *[VII]* implican riesgos no proporcionales puesto que los factores de riesgo 1 y 2 "interactúan" con el tiempo (la variable indicada con cero en nuestra notación). Al modelo *[VII]* se le denomina *saturado* pues contiene tantos parámetros como el número celdas $i_0 \times i_1 \times i_2$ que definen los datos. En la selección de un modelo se busca que éste replique lo más acertadamente posible la informa-

ción contenida en los datos, utilizando para ello el menor número de parámetros posible. Se busca pues que los modelos sean parsimónicos, sobre todo porque la interpretación de modelos simples es relativamente sencilla. Así, si bien los modelos saturados replican exactamente los datos, desde el punto de vista estadístico no son informativos pues lo hacen al costo de utilizar tantos parámetros como datos (reduciendo los grados de libertad a cero).

La teoría de estimación estadística y de prueba de hipótesis de modelos log-lineales como [13] se encuentra descrita en Bishop *et al.* (1975), Fienberg (1977) y Haberman (1978), entre otros. A su vez, Holford (1980) y Laird y Olivier (1981) muestran cómo, con la sola restricción de que las covariables se modelen como variables categoriales, el marco analítico y los procedimientos convencionales de estimación de modelos log-lineales pueden aplicarse para analizar y modelar datos de sobrevivencia. El punto central de sus resultados consiste en probar que la estimación vía máximo de verosimilitud de los parámetros $\underline{B}$ en cualquier modelo de sobrevivencia exponencial o Poisson es equivalente a la estimación, vía máximo de verosimilitud, de modelos log-lineales. En particular, Laird y Olivier utilizan LOGLIN (Olivier y Neff, 1976) un paquete de cómputo basado en el alogaritmo de ajuste interactivo proporcional (AIP), diseñado para estimar modelos log-lineales (Darroch y Ratcliff, 1972) pero que es fácilmente adaptable para estimar modelos de sobrevivencia como [13]. Aitkin y Clayton (1980), utilizando resultados semejantes, describen el uso de otro paquete de cómputo, GLIM (Baker y Nelder, 1978), para estimar modelos de sobrevivencia basados en la distribución exponencial, en la Weibull, y en la distribución de valores extremos. En la sección "Estimación y prueba de hipótesis" mostramos cómo LOGLIN puede utilizarse también para estimar modelos de riesgos en competencia con covariables, utilizando una extensión sugerida por Laird y Olivier (1981).

Modelos de riesgos en competencia con covariables

Notación y definiciones

La teoría de riesgos de competencia se encuentra excelentemente descrita en textos como Chiang (1968, 1980), Birnbaum (1978), y David y Moeschberger (1978). En demografía, las aplicaciones más usuales se refieren a la construcción de tablas de decrementos múltiples

como pueden ser las tablas de mortalidad por causas (Preston, Keyfitz y Shoen, 1972) o bien las tablas de fenómenos "al estado puro", donde se corrige el efecto de fenómenos "perturbadores" (Pressat, 1969).

En esta sección, en lugar de considerar un solo decremento h(t) como en las secciones anteriores, vamos a suponer que los individuos bajo estudio están sujetos a J "causas de muerte" de tal forma que cada una define un riesgo independiente $h_j(t)$, con $j = 1, 2, \ldots, J$.

La teoría correspondiente a este caso está descrita en Kalbfleisch y Prentice (1980), Prentice *et al.* (1978), Holt (1978), Lawless (1982), y Larson (1983), entre otros. Aquí no hacemos más que presentar los aspectos que creemos son más relevantes para el tipo de aplicación que proponemos.

En este caso suponemos que para cada individuo se dispone de una tríada de información (C, T, $\underline{z}$), donde T y $\underline{z}$ son, como antes, el tiempo de sobrevivencia t (que puede ser partido en K intervalos discretos) y un vector de P covariables, y C se refiere a *una* causa de muerte, la j, entre J causas posibles. A nuestras definiciones anteriores estamos añadiendo pues una dimensión más, la correspondiente a la causa de muerte j. Así, las definiciones [1], [2] y [3] resultan ahora:

$$S_j(t) = Pr(T \geq t, C = j), t \geq 0 \qquad [1']$$

$$f_j(t) = -\frac{dS_j(t)}{dt} \qquad [2']$$

$$h_j(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t; C = j \mid T \geq t)}{\Delta t}$$

$$= \frac{f_j(t)}{S(t)} \qquad [3']$$

donde $h_j(t)$ es la función de riesgo específica para la causa j, en presencia (competencia) de todas las demás causas.

La nueva definición [3'] satisface la relación

$$h(t) = \sum_{j=1}^{J} h_j(t)$$

de tal forma que S (t) retiene su sentido original, es decir, como la probabilidad de sobrevivir, hasta el momento t, a las J causas de muerte:

$$S(t)=\exp\left[-\int_o^t h(u)\,du\right]$$

Modelos discretos con covariables

Aquí suponemos que: *a)* la dimensión tiempo-duración de cada función de riesgo $h_j(t)$ puede partirse en K intervalos fijos de tiempo (t_k, t_{k+w_k}) (de duración w_k arbitraria); y *b)* las covariables $\underline{z}$ están medidas categóricamente. Con estos supuestos, el modelo de riesgos proporcionales puede rescribirse como:

$$h(j,t_k|\underline{z})=h_o(j,t_k)\exp \underline{z}\ \underline{B}_j \qquad [14]$$

en donde los componentes son básicamente los mismos que intervienen en [8] y en [10], excepto que ahora tanto la función de base $h_o(j,t_k)$ como los coeficientes de regresión $\underline{B}_j$ se les permite variar arbitrariamente según las J causas de muerte. Las ecuaciones de verosimilitud de [14] así cómo diversos aspectos tocantes a la estimación de los parámetros $\underline{Z}$ se encuentran descritos en Kalbfleisch y Prentice (1980). Larson (1983) muestra a su vez cómo, mediante ciertas adaptaciones, los procedimientos para estimar modelos log-lineales se prestan para estimar modelos de la forma general de [14]. En este trabajo hacemos uso de los resultados de estos autores sin entrar en el detalle de sus derivaciones.

La formulación log-lineal

Para simplificar la notación vamos a suponer que el vector $\underline{z}$ se reduce a una sola covariable, Z,con lo cual las variables que se incluyen en el modelo [14] resultan entonces: la causa de muerte C; el tiempo de sobrevivencia T; y una covariable Z. Los índices de estas variables son, respectivamente, j, k y l. La configuración total de los datos va a estar dada entonces por una tabla de JxKxL dimensiones, donde J, K y L indican el número máximo de categorías de cada variable respectiva.

Con esta notación el modelo saturado correspondiente a [14] puede escribirse como:

$$\log h(j, t_k \mid \underline{Z}) = \log \Theta_{jkl} \qquad [15]$$

$$= U + U_{C(j)} + U_{T(k)} + U_{Z(l)} +$$

$$+ U_{CT(jk)} + U_{CZ(jl)} + U_{TZ(kl)} +$$

$$+ U_{CTZ(jkl)}$$

que no es sino una modificación de [13], donde la primera variable se refiere a la causa de muerte en lugar de a la dimensión tiempo. Como de costumbre, para evitar sobreparametrización, el modelo [15] debe satisfacer las restricciones ANOVA habituales (Bishop *et al.*, 1975; Fienberg, 1977). En lo sucesivo, salvo que la notación lo requiera, vamos a obviar también los subíndices j, k, l.

CUADRO 1
Diversas especificaciones del modelo [15], indicadas por las variables a que se refieren los términos U del modelo (las variables subrayadas constituyen la clase generadora del modelo)

[a]	$\underline{C}$
[b]	$\underline{C}, \underline{T}$
[c]	$\underline{C}, \underline{Z}$
[d]	$\underline{C}, \underline{T}, \underline{Z}$
[e]	$C, T, \underline{CT}$
[f]	$C, Z, \underline{CZ}$
[g]	$C, T, \underline{Z}, \underline{CT}$
[h]	$C, \underline{T}, Z, \underline{CZ}$
[i]	$\underline{C}, T, Z, \underline{TZ}$
[j]	$C, T, Z, \underline{CT}, \underline{CZ}$
[k]	$C, T, Z, \underline{CT}, \underline{TZ}$
[l]	$C, T, Z, \underline{CZ}, \underline{TZ}$
[m]	$C, T, Z, \underline{CT}, \underline{CZ}, \underline{TZ}$
[n]	$C, T, Z\ CT, CZ\ TZ\ \underline{CTZ}$

En el cuadro 1 señalamos todas las posibles especificaciones de modelos jerárquicos que corresponden a eventuales simplificaciones de [15]. Tres clases generales de términos U son identificables en este cuadro: los que se refieren sólo a los riesgos de base (es decir donde no interviene Z); los que se refieren a los efectos de Z considerados

como independientes de T; y los efectos que indican dependencia de Z con respecto a los intervalos de tiempo T. Los modelos [a], [b] y [e] se refieren a la primera clase; en particular [a] especifica sobrevivencia exponencial para cada causa j, mientras que [e] especifica no proporcionalidad de los riesgos de base.

Todos los términos Z y CZ representan efectos proporcionales; los primeros operan para todos los intervalos y para todas las causas, mientras que los segundos son específicos para cada causa. De modo similar, todos los términos T Z y CTZ indican que los efectos de las covariables cambian según los intervalos de tiempo; para todas las causas los primeros, y específicamente por causa los segundos.

Las diferencias que existen entre los modelos del cuadro 1 se refieren al número de términos que se suponen nulos. Por ejemplo, la diferencia entre los modelos [m] y [j] reside en la hipótesis de que $U_{TZ} = 0$, por encima de los términos nulos que especifica [m]: $U_{CTZ} = 0$. (Puesto que [m] contiene a [j], se dice que [j] está anidado en [m] y ambos hacen una jerarquía). La selección de un modelo está orientada pues, además de por ciertos criterios como pueden ser la preferencia por alguna forma funcional particular o consideraciones del diseño muestral, por la prueba de hipótesis sobre la significancia estadística de sus términos, dentro de una jerarquía de modelos anidados. Las dos estrategias más sencillas de selección consisten en ir añadiendo términos al modelo nulo, o bien, ir descartando términos a partir del modelo saturado; en cada paso se juzga la bondad de ajuste del modelo que resulta y la significancia de los términos que se añaden o quitan. No existe sin embargo ninguna "rutina" para la selección del "mejor" modelo pues, en cualquier caso, el proceso de selección depende de la jerarquía de modelos que se escoja. Bishop *et al.* (1975) y Fienberg (1977) tocan en detalle esta dificultad y dan sugerencias de cómo guiar el proceso de selección de un modelo. En nuestro caso, como se verá en la sección "Legalización y terminación de uniones consensuales en México", preferimos partir de un modelo complejo, efectuando la simplificación por pasos según las siguientes categorías: *a)* interacción entre las covariables; *b)* dependencia de las covariables con el tiempo según causas específicas; *c)* dependencia general de las covariables con el tiempo, sin distinción de causas; *d)* asociación de las covariables con el tipo de causa.

Antes damos algunas aclaraciones sobre el modo de estimación.

Estimación y prueba de hipótesis

La estimación de modelos de sobrevivencia mediante métodos log-lineales para el análisis de tablas de contingencia es factible, como dijimos, gracias a la verificación de dos relaciones (Laird y Olivier, 1980): *1)* que los modelos log-lineales del número esperado de casos generados por un proceso Poisson en la realización de una tabla de contingencia son equivalentes a los modelos log-lineales de sobrevivencia como [8] cuando la función de riesgo-base es exponencial y las covariables son discretas; y *2)* que las ecuaciones de verosimilitud de ambos casos (con datos de sobrevivencia de una exponencial escalonada, y datos de una tabla de contingencia Poisson) son equivalentes también. Larson (1983) prueba estas dos relaciones para el caso más general de riesgos en competencia y muestra que los estimadores de máxima verosimilitud de [15] son:

$$\log \Theta_{jkl} = \log \hat{D}_{jkl} - \log E_{kl} \; j = 1, 2, \ldots, J$$

donde E_{kl} es el tiempo de *exposición al riesgo* (sin distinguir la causa o el mecanismo que truncó la exposición) y $\hat{D}_{jkl}$ son los decesos estimados que resultan a su vez de $\hat{D}_{jkl} = E_{kl} \exp$ (modelo log-lineal retenido). Para fines prácticos, la tabla E_{kl} se considera como una constante, y sirve como valor inicial de las D ($D^{o}_{jkl} = E_{kl}$) en el algoritmo AIP mediante el que se estiman los términos U del modelo log-lineal retenido.

Las estimaciones que hacemos aquí se basan en el programa LOGLIN que, como dijimos, utiliza justamente AIP como algoritmo de maximización y permite estimar modelos de sobrevivencia haciendo pivotear la estimación de $\hat{D}_{jkl}$ sobre E_{kl}.

La configuración básica de los datos para utilizar LOGLIN debe ser como sigue, por ejemplo para j = 2, k = 3, y l = 3.

Matriz de casos

		Z_1	Z_2	Z_3
C_1	T_1 T_2 T_3			
C_2	T_1 T_2 T_3			

Matriz de exposición

	Z_1	Z_2	Z_3
T_1 T_2 T_3			
T_1 T_2 T_3			

donde la matriz E_{kl} debe repetirse J veces para igualar la configuración de la matriz D_{jkl}. Por supuesto, cuando se incluye más de una covariable (nuestro ejemplo ha sido trivial a este respecto) hay que añadir las correspondientes dimensiones a la matriz de casos y a la matriz de exposiciones. La última versión disponible de LOGLIN, la versión 1.6, contiene el comando *Read Surv* que permite construir directamente las matrices de casos y exposiciones a partir de un arreglo de datos que guarde ciertas características de formato (Olivier y Neff, 1981).

LOGLIN proporciona también una serie de estadísticos para la prueba de bondad de ajuste de los modelos. En particular proporciona la *razón de verosimilitud* G^2 (y sus correspondientes grados de libertad) que se define como:

$$G^2 = 2\sum 0 - \log (0/E)$$

donde 0 se refiere a las frecuencias observadas en cada celda, y E son los estimadores de máxima verosimilitud de dichas frecuencias. El valor de G^2 se interpreta como la probabilidad, en una distribución χ^2, de que las diferencias entre las frecuencias observadas y las estimadas sean sólo aleatorias, bajo la hipótesis nula de que el modelo sea correcto.

Otro estadístico común de bondad de ajuste es la χ^2 de Pearson (1904)

$$\chi^2 = \sum \frac{(0-E)^2}{E}$$

que se distribuye bajo el supuesto de que la hipótesis nula es correcta, como χ^2, con sus correspondientes grados de libertad.

Una ventaja de G^2 sobre χ^2 es que permite descomponer la prueba de un modelo en dos partes, como se describe ahora. Supongamos que tenemos dos modelos, 1 y 2, con grados de libertad v_1 y v_2, y valores estimados E_1 y E_2, respectivamente. Supongamos también que los términos U del modelo 2 constituyen un subconjunto de los términos del modelo 1, es decir, 1 está anidado en 2. Bishop *et al.* (1975) muestran que, bajo el supuesto de que el modelo 1 es correcto, el cociente de verosimilitud de la diferencia de los modelos, $G^2(2/1) = -2[L(E_2) - L(E_1]$, se distribuye como una X^2 con $V_2 - V_1$ grados de libertad, donde $L(E_1)$ y $L(E_2)$ son las funciones de log-verosi-

militud de los dos modelos, evaluados en $\hat{E}_1$ y $\hat{E}_2$. Si el modelo 1 es el modelo saturado, entonces $G^2(2/1)$ es simplemente la razón de verosimilitud $G^2(2)$ del modelo 2, puesto que en este caso $\hat{E}_1 = 0$. Si el modelo 1 no es el saturado, la prueba $G^2(2/1)$ es entonces una prueba condicional, donde las $\hat{E}_1$ estimadas por 1 operan "como si fuesen" los datos, es decir, como frecuencias observadas.

Por otro lado, Bishop *et al.* (1975) muestran también que una prueba del modelo 2 puede descomponerse en dos partes

$$G^2(2) = G^2(2/1) + G^2(1)$$

la prueba condicional $G^2(2/1)$ que acabamos de describir, y $G^2(1)$ que mide la bondad de ajuste del modelo 1. La primera parte indica qué tan diferentes son las $\hat{E}_2$ del modelo 2 con respecto a las $\hat{E}_1$ del modelo 1. Si las diferencias son pequeñas, los términos U que el modelo 1 añade al 2 son prescindibles; por el contrario, si las diferencias son grandes, todos los términos de 1 son significativos. A la prueba $G^2(2/1)$ se le denomina *condicional* mientras que $G^2(2)$ es una prueba *marginal*

De lo anterior se desprende que, si el modelo 2 se rechaza, ello seguramente se debe al incumplimiento de cualquiera de las dos partes de la prueba. El modo más útil de valerse de estos principios para la selección de un modelo consiste en asegurarse que el modelo 1 se cumpla (es decir, que ajuste satisfactoriamente los datos); si junto a esto, el modelo 2 se acepta, ello implica que los términos que se excluyeron al pasar de 1 a 2 no son significativos. Fienberg sugiere ir excluyendo términos (en una jerarquía de modelos anidados) hasta que la prueba marginal o la prueba condicional sean significativas en un nivel predeterminado (por ejemplo $p \leq 0.05$); el modelo que se selecciona es el último antes de que se cumpla esta condición. En la siguiente sección utilizamos precisamente este modo de proceder en la selección del modelo que nos parece más adecuado para los datos de nuestro ejemplo.

Legalización y terminación de uniones consensuales en México

El ejemplo

Los métodos descritos aquí se ilustran con datos acerca de la legalización y terminación de uniones consensuales provenientes de la Encuesta Mexicana de Fecundidad (EMF), realizada entre julio de 1976 y

marzo de 1977. La EMF incluyó, en el cuestionario individual, una historia de uniones de todas las mujeres entrevistadas. Fueron elegibles para el cuestionario individual todas las mujeres (seleccionadas del cuestionario de hogar) entre 20 y 49 años, y las mujeres entre 15 y 19 años que hubieran tenido algún hijo nacido vivo, o que vivían en alguna forma de unión. El número total de mujeres entrevistadas fue de 7 310 mujeres. En nuestro estudio analizamos la legalización o la separación de 1 316 mujeres viviendo en una primera unión consensual. Las covariables que retuvimos como factores de riesgo son (junto con su nomenclatura y los valores de las categorías):

A = edad de la mujer al inicio de la unión ($\leq 17, > 17$)

E = años de educación (0-1, 2-5, 6 +)

W = si trabaja o ha trabajado desde que se unió (sí, no).

Las otras variables del modelo son:

C = causa de terminación (legalización civil o religiosa o ambas, separación o muerte del compañero)

T = tiempo-duración en años (0-1, 1-2, 2-5, 5-10, 10-15, 15+)

El tiempo total de exposición observado fue 16 692.2 años persona, durante los cuales hubo 655 legalizaciones y 332 separaciones. En el cuadro 2 se señala la ocurrencia de estos eventos según la duración de la unión y según las covariables seleccionadas. Como puede observarse, la mayor parte de las legalizaciones ocurren en el primer año de duración de la unión, mientras que las separaciones están más uniformemente distribuidas entre los 0 y los 10 años de duración. En este cuadro no se indican las "exposiciones al riesgo", pero están contenidas en el análisis que sigue.

La primera parte del análisis consistió en probar las interacciones de segundo y tercer orden entre las covariables A, E y W. Ninguna de éstas fue significativa y no se reportan aquí para no abrumar la presentación con resultados un tanto irrelevantes. El modelo que resulta de ese análisis es el modelo I del cuadro 3, que puede considerarse como el modelo de partida para el resto del análisis.

En seguida probamos si las covariables tienen efectos que dependen de los intervalos de tiempo-duración de un modo específico por causa. En el cuadro 4 se muestra que ninguna de las diferencias entre los modelos 1.1, 1.2 y 1.3 con el modelo I es significativa, con lo cual concluimos que los efectos no varían con la duración específicamente por causa.

La siguiente prueba consiste en inspeccionar la misma dependencia (entre los efectos de los factores y la duración) pero en general, no

CUADRO 2
Legalizaciones y separaciones según las variables retenidas para el análisis

	C_1 = *Legalizaciones*						
T	*A*		*E*			*W*	
	≤ 17	*> 17*	*0-1*	*2-5*	*6+*	*Sí*	*No*
0 - 1	153	144	103	99	95	47	250
1 - 2	45	36	24	40	17	16	65
2 - 5	71	44	44	41	30	22	93
5 - 10	43	41	30	34	20	19	65
10 -15	24	14	20	13	5	10	28
15+	26	14	21	15	4	8	32

	C_2 = *Separaciones*						
T	*A*		*E*			*W*	
	≤ 17	*> 17*	*0 - 1*	*2 -5*	*6+*	*Sí*	*No*
0 -1	32	44	21	35	20	36	40
1 - 2	28	22	14	22	14	15	35
2 - 5	61	35	38	32	26	38	58
5 - 10	33	33	33	17	16	31	35
10 - 15	17	9	15	9	2	14	12
15+	8	10	7	7	4	9	9

Nota: C = causa de terminación; T = tiempo-duración; A = edad a la unión; E = educación; W = trabajo.

específicamente por causa. Estos resultados se encuentran en el cuadro 4 donde se constrastan los modelos II.1, II.2 y II.3 con II. La edad al inicio de la unión parece tener un efecto ligeramente diferencial por duración, pero no es lo suficientemente significativo como para darle mayor consideración. (Una inspección de los términos U señala que las uniones de mujeres que entran en convivencia en edades jóvenes tienen menos propensión a terminar que las uniones de mujeres que comienzan en edades adultas, y ello es más marcado a partir de las duraciones mayores de cinco años.) El resto de los contrastes de las variables incluidas en esta prueba no es significativo.

El siguiente paso del análisis se refiere a dos pruebas sobre los riesgos de base: investigar si son proporcionales, y si son constantes. En el cuadro 4 se aprecia que el término U_{CT} es significativo, es decir, que los riegos de base no son proporcionales. Por su parte, U_T es

altamente significativo, lo cual confirma que los riesgos de base no son constantes (es decir, las terminaciones no siguen una ley exponencial).

En el cuadro 3 se constata que al excluir los términos U_T o U_{CT}, la bondad de ajuste del modelo deja de ser aceptable, lo cual señala que ambos términos deben formar parte del modelo final. De hecho, podría considerarse que el modelo III proporciona una representación adecuada de los datos, sin embargo, su elevado valor p (= .34) invita a seguir excluyendo términos U del modelo.

El siguiente paso del análisis consiste en probar si los efectos de las covariables tienen alguna especificidad por causa, lo cual se logra constrastando los modelos IV.1, IV.2 y IV.3 con IV. En el cuadro 4 podemos ver que la edad a la unión y la educación no afectan diferen-

CUADRO 3
Bondad de ajuste y nivel de significancia de diversos modelos

Modelo	*Clase generadora del modelo*							G^2	*gl*	*p*
I			CTA	CTE	CTW			80.66	84	.58
1.1	CA	TA		CTE	CTW			82.98	89	.66
1.2	CE	TE	CTA		CTW			93.14	94	.50
1.3	CW	TW	CTA	CTE				85.45	89	.59
II	CT	CA	CE	CW	TA	TE	TW	100.55	104	.58
II.1	CT	CA	CE	CW		TE	TW	111.68	109	.41
II.2	CT	CA	CE	CW	TA		TW	113.67	114	.49
II.3	CT	CA	CE	CW	TA	TE		104.86	109	.59
III	CT	CA	CE	CW				130.14	124	.34
III.1 T		CA	CE	CW				181.46	129	.0016
III.2		CA	CE	CW				516.88	134	—
IV= III	CT	CA	CE	CW				130.14	124	.34
IV.1 A	CT		CE	CW				130.18	125	.36
IV.2 E	CT	CA		CW				130.55	126	.37
IV.3 W	CT	CA	CE					193.24	125	—
V A	E	CT	CW					130.62	127	.39
V.1	E	CT	CW					135.07	128	.32
V.2 A		CT	CW					154.03	129	.066

Nota: C = causa de terminación; T = tiempo-duración; A = edad a la unión; E = educación; W = trabajo.

CUADRO 4
Prueba de hipótesis para términos U selectos

Modelos comparados	*Término bajo prueba*	ΔG^2	*gl*	*p*
I.1 - I	CTA	2.32	5	NS
1.2 - I	CTE	12.48	10	NS
1.3 - I	CTW	4.79	5	NS
II.1 - II	TA	11.35	5	S?
II.2 - II	TE	13.12	10	NS
II.3 - II	TW	4.31	5	NS
III.1 - III	CT	51.32	5	S
III.2 - III	T	386.74	10	S
IV.1 - IV	CA	0.04	1	NS
IV.2 - IV	CE	4.41	2	NS
IV.3 - IV	CW	63.10	1	S
V.1 - V	A	4.45	1	NS
V.2 - V	E	23.41	2	S

Nota: C = causa de terminación; T = tiempo-duración; A = edad a la unión; E = educación; W = trabajo.

cialmente el que la unión termine por legalización o por separación, mientras que la condición de actividad sí afecta significativamente el tipo de causa de terminación. El término U_{CW} debe ser incluido pues en el modelo final, mientras que U_{CA} y U_{CE} pueden ser excluidos.

Por último, probamos la importancia de las covariables edad a la unión y educación en forma aislada, por encima de las covariables ya retenidas como significativas. El contraste de los modelos V.1 y V.2 con V indica que la covariable edad a la unión es prescindible, mientras que la covariable educación es significativa y debe ser retenida en el modelo.

Tras este análisis, el modelo finalmente seleccionado es el modelo V.1 que, en términos U, se escribe como

$$\log \Theta = U + U_C + U_T + U_E + U_W + U_{CT} + U_{CW}$$

El modelo V.1 indica, como dijimos: *a)* que los riesgos de base no son constantes ni son proporcionales; *b)* que la condición de actividad afecta el tipo de terminación de la unión, para todas las duracio-

nes; y *c)* que las covariables educación y condición de actividad afectan proporcionalmente los riesgos de terminación de la unión, para las dos formas de terminación, y para todas las duraciones. En particular, el modelo V.1 es un modelo de riesgos proporcionales (en las covariables) como [14].

Los efectos de todos los términos U del modelo V.1 están indicados en el cuadro 5. Algunas interpretaciones que pueden extraerse de este cuadro son las siguientes:

1) Existe mayor propensión a terminar una unión consensual por legalización que por separación (para cualquier estatus en las covariables y para todas las duraciones).

2) El hecho de que la mujer trabaje o haya trabajado está asociado con mayores probabilidades de terminación de la convivencia (para cualquier tipo de terminación y para todas las duraciones).

3) A mayor educación de las mujeres, mayor probabilidad de terminación de la convivencia (para todas las duraciones y para los dos tipos de terminación).

4) La propensión a la terminación de la unión (para ambos riesgos) está concentrada en las primeras duraciones; disminuye paulatinamente con la duración para volver a aumentar en la duración abierta (15 años o más).

5) Las propensiones a legalizar la unión o a separarse no son proporcionales según la duración de la unión. Ocurren mucho más legalizaciones que separaciones en el primer año de duración; entre los 5 y 10 años de duración esta situación se invierte.

6) La condición de actividad de las mujeres influye no sólo en su propensión a terminar la unión, sino también en el tipo de terminación: las mujeres que trabajan tienden a terminar por separación, mientras las mujeres que no trabajan tienden a terminar por legalización.

En términos de su importancia relativa en el modelo (juzgada ya sea por la magnitud de los errores estándar, o por la diferencia de los valores máximo y mínimo en los parámetros de los términos U) las covariables o las interacciones más importantes en el modelo son:

$$U_T > U_{CW} > U_{CT} > U_E > U_C > U_E$$

Es de notar que, aparte de la dimensión tiempo-duración, que es una variable que siempre reviste la mayor importancia en el análisis demográfico, los componentes más informativos del modelo son dos "interacciones": la dependencia del tipo de terminación de la unión

CUADRO 5
Parámetros estimados para el modelo V.1

U_C (1) = .17	U_W (1) = .097	U_E (1) = −.193
(2) = −.17	(2) = −.097	(2) = −.024
		(3) = .216
U_T (1) = .869		
(2) = .246		
(3) = −.068		
(4) = −.346		
(5) = −.582		
(6) = −.117		
U^*_{CT} (11) - (21) = 1.114	U^*_{CW} (11) - (21) = −0.25	
(12) - (22) = 0.244	(12) - (22) = 0.93	
(13) - (23) = −0.136		
(14) - (24) = −0.692		
(15) - (25) = −1.164		
(16) - (26) = −0.234		

* Los parámetros U_{CT} y U_{CW} incluyen la adición de los parámetros marginales U_C, U_T y U_W donde corresponde.

según la condición de actividad de la mujer, y la no proporcionalidad de los riesgos de terminar por legalización o por separación. La variable edad al inicio de la unión, que es una variable que a menudo se utiliza y tiene considerable importancia en el estudio de la dinámica de uniones, parece tener aquí menos relevancia que las variables socioeconómicas que utilizamos en nuestro ejemplo.

Resta por último insistir que este ejemplo no constituye ningún análisis acabado del tema que se propuso; pretende tan sólo ilustrar la implementación de una forma de análisis que creemos que se presta al estudio de un gran número de relaciones demográficas dentro de un esquema de análisis estadístico multivariado.

Bibliografía

Aitkin, M. y D. Clayton (1980), "The Fitting of Exponential, Weibull and Extreme Value Dstributions to Complex Censored Survival Data Using GLIM", *Journal of the Royal Statistical Society*, serie C, núm. 29, pp. 156-163.

Baker, R. J. y J. A. Nelder (1978), *Generalized Linear Interactive Modeling (GLIM)*, Oxford, Numerical Algorithms Group.

Birnmbaum, Z.W (1979), *On the Mathematics of Competing Risks*, Maryland, DHEW (Publicación PHS, 72).

Bishop, Y. M., S. E. Fienberg, y P. W. Holland (1975), *Discrete Multivariate Analysis; Theory and Practice*, Cambridge, MIT Press.

Chiang, C. L (1968), *Introduction to Stochastic Processes in Biostatistics*, Nueva York, Wiley.

——— (1980), *An Introduction to Schocastic Processes and their Apllications*, Nueva York, Krieger.

——— (1984), *The Life Table and its Applications*, Florida, Krieger.

Cox, D. R. (1972), "Regression Models and Life Tables (with Discussion)", *Journal of the Royal Statistical Society*, serie B, núm. 34, pp. 187-220.

——— (1975), "Partial Likelihood", *Biometrika*, vol. 62, pp. 269-276.

Darroch, J. N. y D. Ratcliff (1972), "Generalized Iterative Scaling for Log-Linear Models", *Annals of Mathematical Statistics*, vol. 43, pp. 1470-1480.

David, H. A. y M. L. Moeschberger (1978), *The Theory of Competing Risks*, Nueva York, MacMillan.

Fienberg, S. E. (1977, 1980), *The Analysis of Cross Classified Data*, Cambridge, MIT Press.

Glasser, M. (1967), "Exponential Survival with Covariance", *Journal of the American Statistical Association*, vol. 62, pp. 561-568.

Haberman, S. J. (1978), *Analysis of Qualitative Data*, vols. 1 y 2, Nueva York, Academic Press.

Holford, T. R. (1980), "The Analysis of Rates and Survivorship Using Log-Linear Models", *Biometrics*, vol. 36, pp. 299-305.

Holt, J. D. (1978), "Competing Risks Analysis with Special Reference to Matched Pair Experiments", *Biometrika*, vol. 65, pp. 159-166.

Kalbfleisch, J. D. y R. L. Prentice (1980), *The Statistical Analysis of Failure Time Data*, Nueva York, Willey.

Kaplan, E. L. y P. Meier (1958), "Nonparametric Estimation from Incomplete Observations", *Journal of the American Statistical Association*, vol. 53, pp. 457-481.

Laird, N. y D. Olivier (1981), "Covariance Analysis of Censored Survival Data Using Log-Linear Analysis Techniques", *Journal of the American Statistical Association*, vol. 76, pp. 231-240.

Larson, M. G. (1983), "Covariate Analysis of Competing Risk Data with Log-Linear Models", *Biometrics*, vol. 39.

Lawless, J. F. (1982), *Stastical Models and Methods for Lifetime Data*, Nueva York, Willey.

Menken, Jane *et al.* (1981), "Proportional Hazards Life Table Models: an Illustrative Analysis of Socio-Demographic Influences of Marriage Dissolution in the U.S.", *Demography*, vol. 18, pp. 181-200.

Olivier, D. C. y R. K. Neff (1976), *LOGLIN 1.0 Users Guide*, Boston, Harvard School of Public Health (mimeo.).

——— (1981), *LOGLIN 1.6; Enhancements to the FIT Command*, Boston, Harvard School of Public Health (mimeo.).

Prentice, R. L., *et al.* (1978), "The Analysis of Failure Times in the Presence of Competing Risks", *Biometrics*, vol. 34, pp. 541-554.

Pressat, R. (1969), *L'Analyse Demographique*, Paris, PUF.

Preston, S., N. Keyfitz y R. Shoen (1972), *Causes of Death, Life Tables for National Populations*, Nueva York, Seminar Press.

Schlesselman, J. J. (1982), *Case Control Studies; Design, Conduct, Analisis*, Oxford, Oxford University Press.