QÜESTIIÓ, vol. 25, 2, p. 337-363, 2001

ANÁLISIS DE DURACIÓN MEDIANTE UN MODELO LINEAL GENERALIZADO SEMIPARAMÉTRICO

JESUS ORBE*

Aitkin y Clayton (1980) proponen el análisis de modelos de duración mediante modelos lineales generalizados. En este trabajo extendemos esta metodología permitiendo que el efecto de alguna de las variables explicativas pueda no ser especificado. Así, el modelo propuesto es un modelo lineal generalizado semiparamétrico, con una componente paramétrica donde se especifica la forma funcional concreta del efecto de las variables explicativas sobre la duración, y una componente no paramétrica donde recogemos el efecto de una variable explicativa sin asumir forma funcional alguna. Desarrollaremos el proceso de estimación así como un procedimiento bootstrap para realizar inferencia. Como aplicación, analizaremos con la metodología propuesta el tiempo de supervivencia para una muestra de pacientes diagnosticados de SIDA.

Lifetime data analysis using a semiparametric generalized linear model

Palabras clave: Modelos de duración, censura, bootstrap, modelos lineales generalizados, estimación semiparamétrica

Clasificación AMS (MSC 2000): 62J12, 62N05, 62G09

* Departamento de Econometría y Estadística (E.A. III). Universidad del País Vasco/Euskal Herriko Unibertsitatea. Avenida Lehendakari Aguirre, 83. 48015 Bilbao. E-mail: jo@alcib.bs.ehu.es.

– Recibido en junio de 2000.
– Aceptado en febrero de 2001.

1. INTRODUCCIÓN

Proponemos una extensión al trabajo presentado por Aitkin y Clayton (1980), quienes presentan la posibilidad de estimar una serie de modelos de duración paramétricos utilizando un modelo lineal generalizado para la variable indicador de censura, cuya función de verosimilitud es proporcional a la correspondiente del modelo de duración original. Tomando como base esta idea, extendemos la metodología de estos autores a un conjunto de situaciones más general. Esta extensión permite modelizar aquellas situaciones en las que la forma funcional del efecto de alguna de las variables explicativas sobre la variable de interés es desconocida o simplemente la especificación de una determinada forma funcional nos parece restrictiva. Esta flexibilización se puede realizar de un modo bastante natural basándonos en un modelo lineal generalizado. Así, vamos a extender el trabajo de Aitkin y Clayton a un contexto semiparamétrico. Como ilustración aplicamos la metodología para analizar el efecto de ciertas variables sobre el tiempo de supervivencia, desde el momento del diagnóstico, en una muestra de enfermos diagnosticados de SIDA.

2. CONEXIÓN ENTRE MODELOS DE DURACIÓN Y MODELOS LINEALES GENERALIZADOS

Sea $T_1,\ldots,T_n$ una muestra aleatoria simple para la variable duración, la cual no es observada en su totalidad debido a la existencia de censura [1] y en su lugar observamos

$$Y_i = \min(T_i, C_i), \qquad \delta_i = \begin{cases} 1; & \text{si } T_i \leq C_i \\ 0; & \text{si } T_i > C_i \end{cases},$$

donde $C_1,\ldots,C_n$ son los valores que toma la variable censura C, la cual suponemos independiente de la variable duración T. Además, δ_i es la variable indicador de censura, tomando valor 0 si la observación correspondiente está censurada o valor 1 si no lo está[2].

Supongamos que esa variable duración puede ser explicada con un modelo que pertenece a la clase de modelos de duración con función de riesgo proporcional propuesto por Cox (1972), en el cual se especifica la siguiente modelización para la función de riesgo:

$$\lambda(t;x) = \lambda_0(t)\exp(x^T\beta) = \lambda_0(t)\exp\eta,$$

[1] Situación habitual cuando se analiza esta clase de datos. Si consideramos el caso más habitual, censura por la derecha, tenemos una observación censurada cuando su tiempo de fallo no ha sido observado al finalizar el estudio (para un detallado análisis de los distintos tipos de censura, vease, por ejemplo, Lawless, 1982).

[2] Con esta especificación estamos suponiendo el caso de censura aleatoria, el habitualmente utilizado. Además, añadiremos que este tipo de censura engloba a otros tipos de censura más restrictivos.

donde $\eta = x^T\beta$ es el predictor lineal y $\lambda_0(t)$ la función de riesgo básica. La estimación de los parámetros del modelo puede realizarse maximizando la función de verosimilitud

$$L = \prod_{i=1}^{n} f(y_i, x_i)^{\delta_i} S(y_i, x_i)^{1-\delta_i}. \tag{1}$$

Utilizando las relaciones existentes entre las funciones de supervivencia, riesgo, riesgo acumulado y de densidad, obtenemos las siguientes expresiones:

$$S(t,x) = \exp(-\Lambda_0(t)e^{\eta})$$

$$f(t,x) = \lambda_0(t)\exp(\eta - \Lambda_0(t)e^{\eta}),$$

donde $\Lambda_0(t) = \int_0^t \lambda_0(t)dt$ es la función de riesgo acumulado básica.

Sustituyendo las expresiones anteriores en (1), tomando logaritmos y reordenando términos obtenemos,

$$\ln L = \sum_{i=1}^{n} \delta_i \left[\ln \Lambda_0(y_i) + \eta_i\right] - \Lambda_0(y_i)e^{\eta_i} + \delta_i \ln\left(\frac{\lambda_0(y_i)}{\Lambda_0(y_i)}\right). \tag{2}$$

Tomando $\mu_i = \Lambda_0(y_i)e^{\eta_i}$ tenemos que,

$$\ln L = \underbrace{\sum_{i=1}^{n} (\delta_i \ln \mu_i - \mu_i)}_{(a)} + \underbrace{\sum_{i=1}^{n} \delta_i \ln\left(\frac{\lambda_0(y_i)}{\Lambda_0(y_i)}\right)}_{(b)}. \tag{3}$$

Se puede verificar que el sumando (a) de la expresión anterior es proporcional al logaritmo de la función de verosimilitud correspondiente a una muestra de n variables aleatorias independientes δ_i con distribución de Poisson[3] de media μ_i. Por otra parte, el término (b) no depende de los parámetros β, sólo depende de la función de riesgo básica, la cual puede depender de parámetros de la distribución.

De esta forma, siguiendo el trabajo de Aitkin y Clayton (1980), y dada la función de riesgo acumulado básica $\Lambda_0(t)$, podemos estimar los coeficientes β del modelo tratando a la variable indicadora de censura δ_i como una variable aleatoria con distribución de Poisson de media $\mu_i = \Lambda_0(y_i)e^{\eta_i}$. Es decir, podemos construir un modelo log-lineal de Poisson «auxiliar» al modelo de duración, tal que

$$\ln(\mu_i) = \ln \Lambda_0(y_i) + x_i^T\beta,$$

con una función de verosimilitud proporcional al término (a) de la expresión (3).

[3] $\sum_{i=1}^{n}(\delta_i \ln \mu_i - \mu_i) - \sum_{i=1}^{n} \ln \delta_i!$

El modelo log-lineal de Poisson, es un caso particular de los Modelos lineales generalizados (MLG). Estos modelos son introducidos por primera vez en el trabajo de Nelder y Wedderburn (1972) y pueden considerarse como una generalización del modelo lineal clásico. En este caso concreto tenemos una componente aleatoria con una distribución de Poisson, y una función de enlace logarítmica que nos relaciona a la media con el predictor lineal. Además en este modelo tenemos un término adicional, $\ln \Lambda_0(y_i)$, denominado «offset». Como MLG puede ser maximizado utilizando el procedimiento de estimación habitual en éstos, maximizando la función de verosimilitud o, equivalentemente, aplicando mínimos cuadrados ponderados iterativos (IRLS) (McCullagh y Nelder 1983).

En Aitkin y Clayton (1980) se describe en detalle el proceso de estimación mediante MLG para los modelos de duración con distribución exponencial, Weibull y valor extremo. Por otra parte, Whitehead (1980) propone un procedimiento de estimación análogo a los anteriores pero para aquellos casos en que desconocemos la forma funcional de la función de riesgo básica; es decir, propone la estimación mediante MLG para modelos de función de riesgo proporcional de Cox (1972).

El resto del trabajo se organiza de la siguiente forma. En la Sección 3 mostramos una aplicación utilizando la metodología tradicional en análisis de duración. Posteriormente, y motivándolo en la aplicación anterior, en la Sección 4, presentamos una extensión al trabajo de Aitkin y Clayton (1980) desarrollando el proceso de estimación de este nuevo modelo. En la Sección 5 proponemos un nuevo procedimiento bootstrap para realizar inferencia en el modelo propuesto. Finalmente, la Sección 6 presenta los resultados y conclusiones más importantes.

3. APLICACIÓN

3.1. Datos

Para ilustrar la metodología descrita hemos aplicado ésta para analizar el tiempo de supervivencia desde el momento del diagnóstico, en una muestra compuesta por 461 enfermos diagnosticados de SIDA desde 1984 hasta el comienzo de 1991, residentes en las comunidades autónomas del País Vasco y Navarra. Utilizando la fecha de diagnóstico y la fecha de fallecimiento, o en el caso de las observaciones censuradas, la fecha final del seguimiento (diciembre de 1992), obtenemos la variable de interés, la duración o tiempo de supervivencia desde el momento del diagnóstico medida en número de trimestres. A diferencia de la mayoría de los trabajos realizados en este área, los cuales se han interesado en estudiar la duración del periodo de incubación, nosotros nos hemos centrado en el estudio de la duración de la última etapa. En el desarrollo del virus VIH tenemos tres etapas. La primera de ellas, la conocida como fase «pre-anticuerpos», es la

más corta con una duración de varios meses (aproximadamente el 50% de los enfermos genera anticuerpos antes de los dos meses después de la infección). Esta etapa va desde el momento en que se produce la infección hasta el desarrollo de los anticuerpos o punto de seroconversión, y es el periodo de tiempo donde al enfermo se clasifica como seronegativo. La segunda etapa, etapa de incubación, es la más larga de las tres (aproximadamente la mitad de los infectados desarrollaban la enfermedad antes de los 10 años). Este periodo parte desde el momento de la seroconversión hasta el diagnóstico de SIDA. Durante esta etapa el individuo es clasificado como seropositivo. Y por último, la tercera etapa, que recoge el tiempo de supervivencia desde el diagnóstico del SIDA. El comienzo de esta etapa tiene lugar en el momento en que el individuo desarrolla alguna enfermedad clasificada dentro de las enfermedades relacionadas con el SIDA.

Para ayudar a describir esta variable disponemos de una serie de variables que nos recogen ciertas características de los enfermos. Así, la variable **Edad** recoge la edad del enfermo en el momento del diagnóstico. **Sexo** es una variable ficticia, que toma valor 1 si el enfermo es varón y valor 0 si es mujer. Tenemos información sobre la enfermedad con la cual se le diagnostica el SIDA. Así, la variable **Enfer1** toma valor 1 si la enfermedad de diagnóstico es una infección oportunista, **Enfer2**, si es un linfoma o un sarcoma de Kaposi y **Enfer3** si es debido a una encefalopatía VIH o al síndrome de «agotamiento» VIH. Además, tenemos información sobre la vía de transmisión de la enfermedad: la variable indicador **Sexual** toma valor 1 si la vía de transmisión es sexual, **Drogas** toma valor 1 si la infección se produce por consumo de drogas, **Sanguínea** toma valor 1 cuando el enfermo es infectado por transmisión sanguínea, **Madre-hijo** toma valor 1 si la transmisión se produce de la madre al hijo, y **Otras** cuando se desconoce la vía de transmisión. Por último, la variable **Periodo** es una variable indicador que toma valor 1 cuando la fecha del diagnóstico es posterior a 1987. El motivo de introducir esta variable ficticia es estudiar el posible efecto de la introducción, a mediados de 1987, del fármaco Zidovudine (también conocido como AZT) sobre la supervivencia del enfermo.

3.2. Análisis de duración tradicional

A continuación, analizamos el efecto que tiene cada una de las variables descritas en la sección anterior sobre el tiempo de supervivencia desde el diagnóstico de la enfermedad. Para ello, comenzamos suponiendo que la variable duración sigue una distribución de Weibull, una de las distribuciones más importantes y más utilizadas en la práctica. La distribución de Weibull es lo suficientemente flexible como para englobar distintos tipos de funciones de riesgo (crecientes, decrecientes o constantes en función del valor que tome el parámetro de forma p). Por tanto, comenzamos ajustando un modelo de regresión Weibull.

Tabla 1. Estimación del modelo de regresión Weibull

Variable	Coeficiente	desv. típica	T-ratio	P-valor
Constante	1.6864	0.4265	3.954	0.00007
Sexo	0.0585	0.1285	0.455	0.64913
Periodo	0.2024	0.1029	1.967	0.04915
Enfer1	0.1881	0.2455	0.766	0.44364
Enfer2	−0.0247	0.3057	−0.081	0.93558
Sexual	−0.1829	0.2372	−0.771	0.44070
Drogas	−0.0392	0.2031	−0.193	0.84711
Sanguínea	−0.0114	0.2735	−0.042	0.96671
Madre−hijo	0.4005	0.4429	0.904	0.36593
Edad	−0.0172	0.0065	−2.621	0.00876
σ	0.9899	0.0366	26.99	0.00000

El efecto de las variables X sobre la función de supervivencia y riesgo puede incluirse a través del parámetro de escala λ. Para ello, utilizamos la forma funcional habitual

$$\lambda = e^{-x^T\beta}, \tag{4}$$

donde x^T es el vector (1×10) de valores que toman los regresores, incluyendo la constante, para cada individuo, y β el vector (10×1) de coeficientes asociados a cada regresor. Por tanto, la función de supervivencia quedará especificada como,

$$S(t,x) = \exp[-(e^{-x^T\beta}t)^p], \qquad p > 0, \quad t > 0,$$

y la función de riesgo como

$$\lambda(t,x) = e^{-x^T\beta}p(e^{-x^T\beta}t)^{p-1}, \qquad p > 0, \quad t > 0.$$

Para la especificación (4), este modelo puede reescribirse en términos log-lineales; es decir,

$$\ln(T) = X\beta + \sigma\varepsilon, \qquad \text{donde} \qquad \sigma = p^{-1},$$

y donde ε tiene una distribución valor extremo estándar. La estimación del modelo se realiza maximizando la función de verosimilitud (1). Los resultados de la estimación se muestran en la Tabla 1.

Analizando los resultados de la Tabla 1 podemos apreciar una estimación del parámetro σ igual a 0.9899, prácticamente 1, y además, significativo. Esto nos está indicando que la distribución de la duración puede ser una exponencial. Para contrastar esta hipótesis

podemos construir el estadístico correspondiente al contraste de la razón de verosimilitudes. Es decir; realizamos el siguiente contraste dentro de la clase de modelos de regresión Weibull: $H_o:\ \sigma = 1$ (distribución exponencial) frente a $H_a:\ \sigma \neq 1$ (distribución no exponencial).

Ajustamos los modelos bajo la hipótesis nula y bajo la hipótesis alternativa y calculamos el máximo del logaritmo de la función de verosimilitud para cada caso. Obtenemos unos valores de -713.29, en el modelo exponencial, y -713.25 en el modelo no exponencial.

Si construimos el estadístico tenemos que,

$$\Lambda = 2\{\ln[L(\hat{\beta},\hat{\sigma})] - \ln[L(\tilde{\beta},\sigma = 1)]\} \xrightarrow{\text{d}} \chi^2_1, \tag{5}$$

donde $(\tilde{\beta},\sigma = 1)$ son las estimaciones del modelo restringido, en nuestro caso el modelo exponencial, y $(\hat{\beta},\hat{\sigma})$ son las del modelo general, es decir, del modelo Weibull. Por tanto no encontramos evidencia estadística contraria a la especificación de un modelo de regresión exponencial.

Como consecuencia del contraste, parece razonable ajustar un modelo de distribución exponencial a nuestros datos. Estimamos de nuevo por máxima verosimilitud y obtenemos los resultados recogidos en la Tabla 2.

Si comparamos los resultados de las Tablas 1 y 2, vemos que apenas varían, lo que refuerza la idea de la distribución exponencial.

Podríamos pensar en llevar a cabo un contraste de significación conjunto de todas las variables del modelo, excepto la constante, para estudiar la contribución conjunta de todas las variables sobre el ajuste del modelo. En el caso de que rechazáramos la no significatividad conjunta del modelo, ésta no sería una condición suficiente para considerar al modelo especificado como válido, habríamos de contrastarla con algún contraste de diagnóstico basado en los residuos, lo que realizaremos posteriormente.

Pasamos a realizar el contraste utilizando el estadístico formado por la razón de verosimilitudes. Ajustamos el modelo restrictivo en el que sólo tenemos como variable regresora la constante y obtenemos un valor máximo del logaritmo de la función de verosimilitud de -724.87. Para el modelo menos restrictivo, donde incluimos todas las variables regresoras obtenemos un valor de -713.29. Si calculamos el valor del estadístico para este contraste, que en este caso se distribuye como una χ^2 con 9 grados de libertad, obtenemos un valor de 23.16, superior incluso al cuantil que deja una probabilidad del 1% a su derecha (21.7). Por tanto, podemos concluir que, aún con un nivel de significación del 1%, rechazamos que las variables regresoras en conjunto no contribuyen a la explicación del modelo.

Tabla 2. Estimación del modelo de regresión exponencial

Variable	Coeficiente	desv. típica	T-ratio	P-valor
Constante	1.6844	0.4306	3.912	0.00009
Sexo	0.0577	0.1298	0.445	0.65700
Periodo	0.2049	0.1035	1.980	0.04770
Enfer1	0.1873	0.2479	0.755	0.45000
Enfer2	−0.0257	0.3087	−0.083	0.93400
Sexual	−0.1835	0.2396	−0.766	0.44400
Drogas	−0.0397	0.2052	−0.193	0.84700
Sanguínea	−0.0109	0.2763	−0.040	0.96800
Madre−hijo	0.3982	0.4472	0.890	0.37300
Edad	−0.0172	0.0066	−2.607	0.00913
σ	1	-	-	-

En cuanto al efecto de cada variable, tenemos que el tiempo de supervivencia del individuo se verá afectado por el periodo en el que se le diagnosticó el SIDA, si el diagnóstico del individuo es posterior a 1987, influirá positivamente en su duración. Por tanto, parece que el uso del fármaco zidovudine, más conocido como AZT, alarga el tiempo de supervivencia y reduce el riesgo, obteniendo unos tiempos de supervivencia, a partir del diagnóstico, superiores.

En cuanto a la variable edad, también parece ser relevante para explicar el tiempo de supervivencia del individuo. A mayor edad, menor será el tiempo de supervivencia y, por tanto, mayor el riesgo.

Una vez tenido en cuenta el efecto de estas variables sobre el tiempo de supervivencia, parece que variables como sexo, el tipo de enfermedad con la que se le diagnostica el SIDA o la categoría de transmisión a la que pertenece no influyen significativamente sobre el tiempo que sobrevive el enfermo.

Además, para los modelos de regresión exponencial y para la especificación $\lambda(x,\beta) = \exp(x^T\beta)$, es posible interpretar los coeficientes en términos de la duración media. La media de una variable aleatoria con distribución exponencial y función de densidad $f(t) = \lambda e^{-\lambda t}$, es $1/\lambda$. En nuestro modelo habíamos especificado $\lambda(x,\beta) = e^{-x^T\beta}$, entonces $1/\lambda = e^{x^T\beta}$. Si tomamos logaritmos tenemos que ln(duración media)$= x^T\beta$, de donde

$$\frac{\partial \ln(\text{duración media})}{\partial x} = \beta.$$

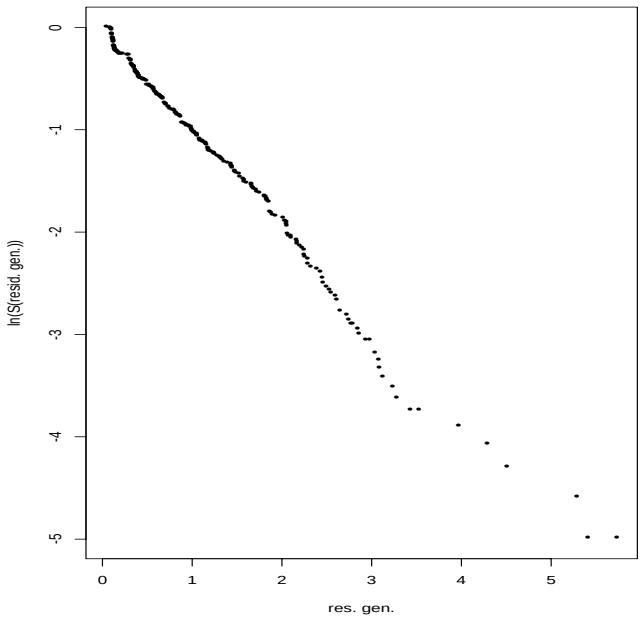

Figura 1. Contraste de diagnóstico

Por tanto, β recogerá la variación porcentual de la duración media ante variaciones de la variable regresora *x*. En el caso de nuestro estudio, esta interpretación sólo tiene sentido para la variable edad (el resto son variables ficticias). Como $\beta_{10} = -0.0172$, este valor nos indica que un aumento de un año en la edad del individuo en el momento de diagnóstico provocará un descenso del 1.7% en el tiempo de supervivencia medio.

Para que los contrastes y conclusiones sobre las estimaciones tengan validez, tenemos que asegurarnos de que el modelo de regresión exponencial se ajusta a nuestros datos realizando un contraste de diagnóstico. Un contraste de diagnóstico sencillo y frecuentemente utilizado es el basado en los residuos generalizados ($\hat{\Lambda}(y_i, x_i)$) o las estimaciones de la función de riesgo integrado[4].

Comenzaremos con un contraste gráfico, donde representamos el logaritmo de la función de supervivencia estimada para los residuos generalizados, mediante el método de

[4]Los residuos generalizados $\hat{\Lambda}(y_i, x_i)$ tienen una distribución exponencial estandar bajo la hipótesis nula de especificación correcta del modelo.

Kaplan-Meier, frente a los residuos generalizados. Bajo la hipótesis nula de especificación correcta, debemos obtener una línea recta que pase por el origen, y de pendiente menos uno. De la Figura 1 podemos suponer que la especificación es aproximadamente la correcta.

Para corroborar el contraste gráfico, y basandonos en la misma idea, realizamos un contraste más formal utilizando el estadístico propuesto por Kiefer (1988)

$$\frac{[\sum_{i=1}^{n} \hat{\Lambda}(y_i, x_i)^2] - 2n}{\sqrt{20n}},$$

con distribución asintótica $N(0,1)$, bajo la hipótesis de correcta especificación del modelo. Antes de computar el estadístico tenemos que ajustar las observaciones censuradas, sumándoles el valor medio[5], en este caso un 1. El valor del estadístico es de 0.13. Por tanto, no encontramos evidencia contraria a la especificación de un modelo de regresión exponencial.

Para finalizar esta sección señalar que, como es lógico, se obtienen las mismas estimaciones utilizando un MLG auxiliar siguiendo la propuesta de Aitkin y Clayton (1980).

4. ANÁLISIS DE DURACIÓN MEDIANTE UN MLG SEMIPARAMÉTRICO

Si analizamos el modelo ajustado en la sección previa nos encontramos un modelo donde el efecto de las variables explicativas es introducido de una forma paramétrica. Es decir, estamos imponiendo una determinada relación (lineal) entre éstas y el logaritmo de la función de riesgo, en el caso de expresar el modelo como caso particular de los modelos de riesgo proporcional, o con el logaritmo de la duración, en el caso de especificar un modelo log-lineal. En algunas situaciones podríamos considerar que la relación paramétrica especificada para alguna de las variables explicativas es muy restrictiva, pudiendo resultar más adecuado introducir el efecto de esta variable de una forma no paramétrica. Así, tendríamos una especificación semiparamétrica, con una parte en la que se recogen variables relacionadas de una forma lineal con la variable a explicar, y otra parte, en la que no se especifique una particular dependencia paramétrica sobre la variable a analizar. Es decir, permitiríamos que los datos reflejaran esta relación mediante una curva de suavizado no paramétrica. Con esta generalización o extensión ampliamos de forma considerable el campo de aplicación de la metodología anterior. Esta extensión nos permite modelizar situaciones en las que no conocemos la forma funcional del efecto de una variable explicativa sobre la variable a explicar, o situaciones en las que suponer una dependencia lineal, u otra cualquiera, entre alguna

[5]Para más detalles sobre este contraste consultar Kiefer (1988).

de las variables explicativas y la variable a analizar sea un supuesto bastante fuerte, o incluso carezca de sentido.

Por tanto, como se puede apreciar la propuesta que vamos a presentar a continuación podría aplicarse en un importante número de situaciones. Un ejemplo ilustrativo del tipo de situaciones que podrían estimarse bajo esta propuesta se recoge en la Sección 3.

En el modelo de la Sección 3 la variable explicativa periodo intenta recoger el efecto de la introducción, a mediados del año 1987, del fármaco zidovudine. Esta variable está construida como una variable ficticia que toma dos valores: valor 1 indicándonos que el diagnóstico del individuo se ha producido con posterioridad a 1987, y valor 0 en caso contrario. Resulta bastante restrictivo dividir el efecto periodo de diagnóstico en dos grupos (antes y después de 1987). Además, parece más lógico o adecuado suponer que el efecto no va ser tan brusco como queda especificado por esa variable ficticia. Por tanto, en esta sección introducimos una componente adicional en el modelo compuesta por una función que depende del periodo de diagnóstico y no especificamos su forma funcional. Así podemos recoger el efecto que tratábamos de recoger, ahora de una forma gradual, además de la evolución completa del efecto que tiene el periodo en que se le diagnostica la enfermedad sobre la supervivencia del individuo.

Por tanto, la extensión que estamos proponiendo al trabajo de Aitkin y Clayton (1980) consiste en considerar modelos de duración con la siguiente función de riesgo

$$\lambda(t;x) = \lambda_0(t)\exp(x^T\beta + h(r)), \tag{6}$$

donde $h(r)$ es una función sin especificar con la cual se recoge el efecto de la variable explicativa R. Así, hemos pasado de un predictor lineal paramétrico $\eta = x^T\beta$ a uno semiparamétrico $\eta = x^T\beta + h(r)$.

Para la estimación de este modelo, al igual que para su equivalente paramétrico, proponemos maximizar la función de verosimilitud, aunque, en este caso consideramos una función de verosimilitud penalizada. Es decir, introducimos un término adicional que penaliza la no suavidad de la función h, y cuyo objetivo es hacer identificable la estimación de β. Por lo tanto, en este proceso de estimación queremos un buen ajuste y una función h lo más suave posible. Consideramos como funciones candidatas a todas las funciones pertenecientes al espacio de Sobolev de orden m $(W_2^m[a,b])$; es decir, todas aquellas funciones cuya derivada m-ésima al cuadrado es integrable en el intervalo $[a,b]$.

La bondad de ajuste dependerá del criterio de optimización elegido. Para el caso de la estimación por máxima verosimilitud, éste estará recogido en la función de verosimilitud.

La medida de suavidad para funciones $h \in (W_2^m[a,b])$ puede estar recogida por $\int_a^b \left[h^{(m)}(r)\right]^2 dr$. En la práctica lo habitual es considerar el caso $m = 2$.

Así, la estimación del modelo puede realizarse mediante la función de verosimilitud penalizada (Good y Gaskins, 1971), la cual considera o tiene en cuenta estas dos características. Para el modelo que estamos considerando, modelo (6), el logaritmo de la función de verosimilitud penalizada tiene la siguiente expresión,

$$\Pi = \sum_{i=1}^{n} \left\{ \delta_i \left[\ln \Lambda_0(y_i) + x_i^T \beta + h(r_i) \right] - \Lambda_0(y_i) e^{x_i^T \beta + h(r_i)} + \delta_i \ln \left(\frac{\lambda_0(y_i)}{\Lambda_0(y_i)} \right) \right\} -$$

$$- \frac{1}{2} \alpha \int_a^b [h''(r)]^2 \, dr. \tag{7}$$

El parámetro de suavizado α, refleja la importancia que damos a la suavidad de la función y a la bondad del ajuste del modelo. Para un valor de α grande estamos dando más importancia a la suavidad, penalizando fuertemente las funciones estimadoras con segunda derivada elevada. Para un valor pequeño estamos dando mayor importancia al buen ajuste del modelo.

De forma análoga al tipo de modelo considerado en la Sección 2, la estimación puede realizarse construyendo un MLG auxiliar, en este caso semiparamétrico, con un logaritmo de la función de verosimilitud proporcional a (7).

El MLG semiparamétrico, auxiliar a este modelo de duración concreto, es un modelo log-lineal de Poisson para la variable indicador de censura δ, con una función de enlace logarítmica y el siguiente predictor lineal:

$$\ln \mu_i = \ln \Lambda_0(y_i) + x_i^T \beta + h(r_i)$$

El logaritmo de la función de verosimilitud penalizada de este modelo auxiliar viene dado por:

$$\Pi = \sum_{i=1}^{n} \delta_i \ln \mu_i - \mu_i - \sum_{i=1}^{n} \ln \delta_i! - \frac{1}{2} \alpha \int_a^b \left[h''(r) \right]^2 dr. \tag{8}$$

Si se sustituye μ_i por su valor $e^{\ln \Lambda_0(y_i) + x_i^T \beta + h(r_i)}$, se puede comprobar que esta expresión es proporcional a (7). Por tanto, las estimaciones en una y otra expresión son las mismas.

Antes de pasar a maximizar la expresión (7), señalaremos que se puede demostrar que la solución, para la función h, al problema de maximizar (7) es una función «spline» cúbica natural. De esta forma y utilizando las propiedades de este tipo de funciones podemos reexpresar (7) (y de forma equivalente (8)) como

$$\Pi = \sum_{i=1}^{n} \left\{ \delta_i \left[\ln \Lambda_0(y_i) + x_i^T \beta + (Nh)_i \right] - \Lambda_0(y_i) e^{x_i^T \beta + (Nh)_i} + \delta_i \ln \left(\frac{\lambda_0(y_i)}{\Lambda_0(y_i)} \right) \right\} -$$

$$- \frac{1}{2} \alpha h^T K h, \tag{9}$$

donde ahora h es el vector de valores $h_j = h(r_j)$, para $j = 1,\ldots,d$ donde d indica el número de valores distintos que toma la variable R, la matriz N se conoce como la matriz incidencia y su función consiste en asignar a cada elemento el valor que le corresponde de la variable que hemos introducido de forma no paramétrica y K es una matriz que se construye utilizando ciertas propiedades de las funciones spline cúbicas naturales[6].

Para maximizar respecto a los coeficientes β y a h podemos utilizar el algoritmo de Fisher scoring. La aplicación de este algoritmo, como se puede demostrar (la demostración en detalle puede encontrarse en Orbe, 2000, pag. 144-146), es equivalente a la resolución del siguiente sistema de ecuaciones simultáneas

$$
\begin{aligned}
X^T W X\beta &= X^T W (Z^* - Nh) \qquad (a)\\
(N^T W N + \alpha K)h &= N^T W (Z^* - X\beta) \qquad (b)
\end{aligned}
\tag{10}
$$

donde el elemento i-ésimo del vector Z^* es:

$$
z_i^* = \ln \Lambda_0(y_i) + x_i^T \beta + (Nh)_i + (\delta_i - \mu_i)\frac{1}{\mu_i} \tag{11}
$$

donde $\mu_i = \Lambda_0(y_i)e^{x_i^T\beta+(Nh)_i}$, W es una matriz de ponderaciones donde los elementos de la diagonal principal son de la forma[7]

$$
w_{ii} = \Lambda_0(y_i)e^{x_i^T\beta+(Nh)_i}, \tag{12}
$$

Para obtener las estimaciones de β y h podemos realizar un procedimiento de «backfitting» (Buja, Hastie y Tibshirani, 1989) entre las ecuaciones (10a) y (10b), hasta alcanzar la convergencia. Así, por una parte, si en la ecuación (10a) conocemos h, el vector de coeficientes β se obtiene regresando por mínimos cuadrados ponderados las diferencias $(Z^* - Nh)$ sobre la matriz de variables regresoras X (las variables de la componente paramétrica). Las ponderaciones serán las anteriormente indicadas. Por otra parte, si conocemos β en (10b) podemos obtener h mediante un suavizador spline cúbico natural aplicado a las diferencias $(Z^* - X\beta)$.

Resumiendo, el procedimiento de estimación completo comienza con la construcción de la matriz incidencia N. Posteriormente, iniciamos el proceso iterativo tomando $\hat{\beta} = 0$, calculamos las estimaciones iniciales del vector h regresando por mínimos cuadrados ordinarios el logaritmo de la variable indicadora de censura $\ln\delta$ sobre la matriz de incidencia. Es decir, $\hat{h} = (N^T N)^{-1} N^T (\ln(\delta))$. Con estas dos estimaciones iniciales, construimos la estimación inicial de $\hat{\mu}_i = \Lambda_0(y_i)e^{x_i^T\hat{\beta}+(N\hat{h})_i}$ y, aplicando la función de enlace

[6] Para más detalles véase, por ejemplo, Green y Silverman (1994), Cap. 2.

[7] Como es lógico μ_i y w_{ii} coinciden, puesto que, por una parte, tenemos una función enlace logarítmica y, por otra parte, en una distribución de Poisson la media y varianza coinciden.

logarítmica, obtenemos el valor inicial del predictor lineal $\hat{\eta}_i = \ln \Lambda_0(y_i) + x_i^T \hat{\beta} + (N\hat{h})_i$. Utilizando las estimaciones de μ y η, obtenemos el vector Z^* siguiendo la expresión (11) y la matriz de ponderaciones W siguiendo la expresión (12). Una vez obtenidos estos valores iniciales, comenzamos con el procedimiento de backfitting, sustituyendo de forma alternativa las estimaciones de (10a) y (10b) hasta que se produzca la convergencia.

En nuestro caso hemos visto que es adecuado proponer una distribución exponencial para la variable T. Por lo tanto, en la expresión (9) sustituimos las expresiones de las funciones de riesgo y riesgo acumulado por las correspondientes de una variable con distribución exponencial. Así, $\Lambda_0(t) = t$ y $\lambda_0(t)/\Lambda_0(t) = 1/t$. Y para maximizar el sistema (10) previamente debemos de realizar la misma sustitución en las expresiones (11) y (12).

En cuanto a la elección del parámetro de suavizado, existen dos aproximaciones al problema diferentes. Por una parte, una aproximación subjetiva que comtempla la posibilidad de que este parámetro sea libremente escogido por el investigador. Esta aproximación es la más utilizada en la práctica. Por otra parte, tenemos una alternativa automática donde el parámetro de suavizado es elegido por los datos. Entre estos criterios automáticos quizá el método más conocido sea el de validación cruzada. Una posibilidad interesante (realizada en este trabajo) sería la de combinar ambas aproximaciones. Como punto de partida utilizaremos un criterio automático como el de validación cruzada generalizada y posteriormente estimaremos el modelo con otros valores.

5. ANÁLISIS DE LAS ESTIMACIONES

Una vez estimados los parámetros del modelo y la función no paramétrica, se nos presenta el problema del análisis de la significatividad o, en general, de realizar inferencia. Dada la componente no paramétrica, podríamos pensar en utilizar contrastes asintóticos (Hastie y Tibshirani, 1990). En lugar de utilizar este tipo de contrastes hemos optado por realizar el estudio de las estimaciones obtenidas mediante técnicas bootstrap. Una de las ventajas que presenta el bootstrap es la posibilidad de analizar las propiedades y realizar inferencia incluso con tamaños de muestra reducidos. Sin embargo, no existe un método bootstrap específico adaptable al modelo propuesto, por lo que procedemos a la elaboración de uno.

Aplicaremos un bootstrap en regresión, ya que disponemos de un conjunto de observaciones no homogéneas, donde la heterogeneidad la tratamos de recoger a través de una serie de variables explicativas utilizando un modelo de regresión. Además, al suponer una distribución concreta para la variable de interés, desarrollaremos un bootstrap paramétrico.

La idea del bootstrap en regresión es la misma que la del bootstrap para modelos homogéneos. Dado que el modelo que estamos considerando parece el adecuado para nuestros datos, realizamos un bootstrap en regresión basado en el modelo. Este procedimiento consiste en obtener la remuestra bootstrap para la perturbación del modelo y, siguiendo la especificación del modelo, construir la remuestra bootstrap para la variable respuesta (para más detalles sobre las técnicas bootstrap, véase, por ejemplo, Efron y Tibshirani, 1993, y Davison y Hinkley, 1997).

Por otra parte, dado que en la muestra tenemos observaciones censuradas y esto tiene que reflejarse en las remuestras bootstrap, tenemos que aplicar un bootstrap adecuado para datos censurados. Tenemos dos posibilidades de remuestreo en el caso de muestras con censura. Efron (1981) propone estimar las funciones de distribución Kaplan-Meier (Kaplan y Meier, 1958) para la variable de interés $\hat{F}_n$ y lo mismo para la variable censura $\hat{G}_n$, posteriormente generar con ambas funciones de distribución sendas muestras para la variable de interés $t_1^*, \ldots, t_n^*$ y para la variable censura $c_1^*, \ldots, c_n^*$, y considerar la siguiente remuestra bootstrap,

$$y_i^* = \min(t_i^*, c_i^*), \qquad \delta_i^* = \begin{cases} 1; & \text{si } t_i^* \leq c_i^* \\ 0; & \text{si } t_i^* > c_i^* \end{cases} .$$

La otra posibilidad, presentada por Reid (1981), consiste en tomar una muestra de observaciones independientes e idénticamente distribuidas con la función de distribución, estimada mediante el estimador Kaplan-Meier, de la variable de interés y considerar la correspondiente función de distribución empírica.

Akritas (1986) demuestra que el plan de remuestreo de Efron es mejor que el de Reid. Además, para el caso de censura aleatoria, Efron demuestra que realizar lo anterior es equivalente a remuestrear con reemplazamiento sobre los pares de variable observada e indicador de censura $(y_1, \delta_1), \ldots, (y_n, \delta_n)$.

Hay que señalar que estos dos procedimientos de generación de muestras bootstrap, para muestras con observaciones censuradas, están pensados para el caso de muestras homogéneas; es decir, para situaciones en las que no tenemos variables explicativas que influyen sobre la variable a explicar, y para el caso en que desconocemos las funciones de distribución de la variable de interés y de la variable censura. Sin embargo éste no es nuestro caso. En nuestro problema, estamos suponiendo una distribución para la variable de interés T (distribución exponencial) y no estamos suponiendo distribución alguna para la variable censura C y, además, tenemos variables explicativas en el modelo. Por lo tanto, para solucionar este problema, tenemos que proponer un nuevo procedimiento generador de muestras bootstrap, adecuado a los supuestos del modelo. Hay que señalar que la propuesta de Efron (para el caso de no suponer la distribución de la variable duración y la variable censura), aún seguiría siendo valida para el caso heterogéneo siempre y cuando supongamos que la variable censura siga el mismo modelo de regresión propuesto para la variable duración.

El modelo concreto que estamos considerando es un modelo de regresión exponencial semiparamétrico, es decir, tenemos la siguiente función de densidad para la variable de interés T:

$$f(t;x) = \lambda e^{-\lambda t}; \quad \text{donde} \quad \lambda = e^{-(x^T\beta + h(r))}.$$

Para la variable censura C no estamos suponiendo distribución alguna.

Para este tipo de modelos proponemos el siguiente procedimiento para generar las remuestras bootstrap:

Paso 1: Ajustar el modelo (6) para el caso de una distribución exponencial.

Paso 2: Generar las perturbaciones bootstrap $\varepsilon_1^*, \ldots, \varepsilon_n^*$, con una distribución valor extremo mínimo.

Paso 3: Obtener la muestra bootstrap para la variable de interés basándonos en el modelo

$$\ln T_i^* = x_i^T \hat{\beta} + \hat{h}(r_i) + \varepsilon_i^*; \quad \text{para} \quad i = 1, \ldots, n.$$

Paso 4: Obtener la muestra bootstrap para variable censura generando una muestra de n observaciones a partir de la función de distribución G de la variable censura.

Paso 5: Comparando las remuestras bootstrap para la variable de interés (paso 3) y la variable censura (paso 4), obtenemos la variable observada bootstrap Y^*, y la correspondiente variable indicador bootstrap δ^*,

$$y_i^* = \min\{t_i^*, c_i^*\}, \qquad \text{y} \qquad \delta_i^* = \begin{cases} 1; & \text{si } t_i^* \leq c_i^* \\ 0; & \text{si } t_i^* > c_i^* \end{cases}.$$

Paso 6: Estimar el modelo (6) (para el caso exponencial) utilizando la información disponible en la remuestra bootstrap.

Paso 7: Volver al paso 2 y repetir el proceso M veces.

Para obtener las estimaciones del modelo (6), en el paso 1, desarrollamos el proceso de estimación propuesto en la sección anterior para el caso particular de una distribución exponencial para la variable T. En el paso 2 estamos considerando el modelo lineal para el logaritmo de la duración[8]. Por lo tanto, en este paso, obtenemos la remuestra

[8] Como suponemos una distribución exponencial de parámetro $\lambda = e^{-(x^T\beta + h(r))}$ para la variable duración, al tomar la transformación logarítmica podemos reescribir el modelo en términos log-lineales como $\ln T = X\beta + h(r) + \varepsilon$ donde, entonces, ε tiene una distribución valor extremo mínimo.

bootstrap de las perturbaciones realizando un bootstrap paramétrico, donde consideramos una distribución valor extremo para las perturbaciones. En el paso 3, como acabamos de comentar, utilizamos la expresión log-lineal de nuestro modelo (6) (ver pie de página 8) para obtener la remuestra bootstrap de la variable de interés. En el paso 4, generamos la variable censura sin considerar ningun supuesto adicional al modelo, que contemple una relación determinada entre las variables explicativas y ésta. La función de distribución G de la variable censura es desconocida y la estimamos utilizando el estimador de Kaplan-Meier, $\hat{G}_n$, adecuado para esta variable. En el paso 6, y como en el paso 1, utilizamos el procedimiento de estimación descrito en la Sección 4. Por último, indicaremos que el número de remuestras bootstrap a considerar depende del objetivo del estudio, si únicamente deseamos calcular las desviaciones típicas de las estimaciones obtenidas, un valor de $M = 200$ puede ser suficiente para obtener unos valores fiables. En cambio, si nuestro objetivo es más ambicioso, y deseamos construir intervalos de confianza, tenemos que considerar un número sensiblemente superior (al menos $M = 1000$), para tener una buena estimación de los percentiles en las colas de la distribución.

6. RESULTADOS Y CONCLUSIONES

Como ilustración de las dos secciones anteriores aplicamos la metodología descrita al conjunto de datos presentados en la Sección 3. Así, la motivación de la extensión del modelo paramétrico, ajustado en la Sección 3, a uno semiparamétrico, tiene su fundamento en el supuesto más razonable de un efecto real, de la introducción del fármaco AZT, más suave o más gradual que el especificado en la Sección 3, utilizando una variable ficticia. Por lo tanto, ahora consideramos un modelo semiparamétrico, concretamente ajustamos el modelo (6) para el caso particular de una variable T con distribución exponencial. El efecto de las variables explicativas quedará dividido en dos términos. El paramétrico, donde recogemos el efecto de todas las variables explicativas excepto la variable periodo de diagnóstico[9] cuyo efecto sera recogido de una forma no paramétrica a través de una función h. El parámetro de suavizado toma un valor igual a 50.

Los resultados de esta estimación y del posterior análisis, de las estimaciones obtenidas mediante técnicas bootstrap, son presentados en las Tablas 3, 4 y Figura 2. Indicaremos que el número de remuestras bootstrap considerado para este análisis es de $M = 1999$.

[9]Ahora, a diferencia de la Sección 3, la variable periodo no va a ser una variable ficticia. En su lugar vamos a crear una nueva variable periodo de diagnóstico que toma valor 1 para los individuos diagnosticados de SIDA en el primer trimestre de la muestra, valor 2, para los diagnosticados en el segundo trimestre, y así sucesivamente, hasta el último trimestre de diagnóstico existente en la muestra.

Tabla 3. Estimaciones de los coeficientes β de la componente paramétrica

Variable	Coeficiente	Desv. típica
Constante	1.37063	0.40330
Sexo	0.02259	0.13103
Enfer1	0.19599	0.25522
Enfer2	0.13172	0.31948
Sexual	−0.23432	0.24722
Drogas	−0.10928	0.20709
Sanguínea	−0.00008	0.28563
Madre-hijo	0.17904	0.47801
Edad	−0.01870	0.00643

Tabla 4. Intervalos de confianza al 95% para las estimaciones de los coeficientes β

	Intervalos bootstrap		Intervalos asintóticos	
Variable	**Lim. Inf.**	**Lim. Sup.**	**Lim. Inf.**	**Lim. Sup.**
Constante	0.5391	2.1134	0.5171	2.2241
Sexo	−0.2176	0.2659	−0.2342	0.2794
Enfer1	−0.2796	0.7088	−0.2893	0.6813
Enfer2	−0.4584	0.7571	−0.4907	0.7542
Sexual	−0.6794	0.2768	−0.7090	0.2404
Drogas	−0.4636	0.3413	−0.5143	0.2957
Sanguínea	−0.5258	0.5947	−0.5451	0.5449
Madre−hijo	−0.6230	1.1253	−0.7045	1.0626
Edad	−0.0308	−0.0056	−0.0317	−0.0056

La Tabla 3 muestra la estimación de los coeficientes β para aquellas variables introducidas en la componente paramétrica del modelo junto a la estimación bootstrap de sus desviaciones típicas. La Tabla 4, además de presentar los intervalos de confianza bootstrap percentil BC (al 95%) para estos coeficientes β, también incluye los intervalos de confianza asintóticos, que habitualmente se calculan en el contexto de los modelos aditivos generalizados[10]. Los resultados son similares, aunque en líneas generales se puede observar que la amplitud de los intervalos bootstrap es menor. Además, estos

[10]Para más detalles veáse Hastie y Tibshirani (1990).

son válidos incluso para muestras de tamaño reducido. La Figura 2 nos presenta la estimación de la componente no paramétrica, la función $h(r)$, así como, las bandas de confianza bootstrap percentil al 95% (para una detallada descripción sobre intervalos de confianza bootstrap ver, por ejemplo, Efron, 1987 y Efron y Tibshirani, 1986).

Antes de pasar a interpretar los resultados obtenidos tenemos que señalar que las estimaciones presentadas en las tablas y figura mencionadas anteriormente indican el efecto de esas variables sobre el logaritmo de la duración. El efecto sobre la función de riesgo va a ser el mismo pero de signo contrario al presentado en las tablas y figura. Una vez aclarada esta cuestión pasamos a reseñar los resultados más relevantes.

En cuanto a las variables introducidas en la componente paramétrica del modelo, indicaremos que únicamente la variable edad resulta significativa para explicar el tiempo de supervivencia del enfermo. A mayor edad en el momento del diagnóstico tenemos un tiempo de supervivencia menor para el enfermo. El resto de las variables de la componente paramétrica resultan no significativas para explicar la supervivencia. Estos mismos resultados se han obtenido en otros trabajos como se recoge en la síntesis de resultados, obtenidos por diferentes autores aplicando diferentes metodologías, presentada en Brookmeyer y Gail (1993).

En cuanto a la componente no paramétrica, propuesta para flexibilizar la más restrictiva aproximación realizada en la Sección 3 (donde se dividía el periodo de estudio en dos partes mediante una variable ficticia), podemos apreciar además del efecto de la introducción del fármaco AZT, la evolución del efecto del periodo de diagnóstico sobre el tiempo de supervivencia. Así, podemos observar una tendencia ligeramente creciente, mayores tiempos de supervivencia, a medida que nos desplazamos de los primeros periodos de diagnóstico. Esta suave tendencia creciente puede venir provocada por el cada vez mayor conocimiento de la enfermedad con el paso del tiempo, lo cual puede originar diagnósticos cada vez más precoces, aumentando así el tiempo de supervivencia desde el momento del diagnóstico. Posteriormente, observamos una fuerte aceleración, en este efecto positivo, sobre la supervivencia, para finalmente mantenerse en niveles máximos. Aquí habría que recordar que la introducción del AZT se produce a mediados de 1987 (alrededor del trimestre 13).

Por lo tanto, la Figura 2 parece mostrarnos un efecto beneficioso de la introducción del fármaco, provocando una importante mejora en el tiempo de supervivencia del enfermo. Como se puede apreciar en la figura, la aceleración de este efecto positivo se produce varios trimestres antes de la introducción del fármaco, lo cual resulta bastante lógico, ya que individuos diagnósticados de SIDA, antes de la introducción del fármaco, también van a recibir el fármaco (aunque no desde un principio) y por tanto, también se benefician de los resultados positivos de éste. Señalaremos que este efecto positivo del AZT también se obtiene en el modelo de la Sección 3 y en otros trabajos como se señala en Brookmeyer y Gail (1993). Entre ellos podemos citar, por ejemplo, Lemp y otros (1990) y Moore y otros (1991). Sin embargo, tenemos que añadir que con la especi-

ficación semiparamétrica que proponemos en la Sección 4 somos capaces de capturar el efecto del AZT de una forma gradual y más flexible, cosa que no podemos hacer bajo una especificación con variable ficticia, puesto que esta especificación está considerando un efecto repentino o brusco. Además, la especificación semiparamétrica nos permite, aparte del efecto de la introducción del AZT, analizar la evolución total del efecto periodo de diagnóstico sobre la supervivencia.

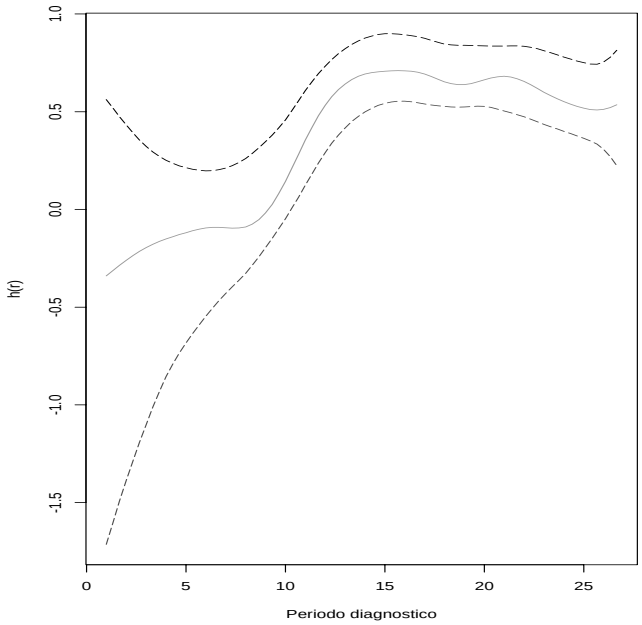

Figura 2. Estimación e intervalo de confianza bootstrap (95%) para la componente no paramétrica

Antes de finalizar, indicaremos que la principal motivación del trabajo presentado ha sido la propuesta de extensión del trabajo de Aitkin y Clayton (1980). Por tanto, el análisis empírico llevado a cabo, prentende, básicamente, ilustrar la metodología propuesta. Aún así, se han extraido una serie de resultados interesantes que, además, nos pueden ayudar a entender mejor la relevancia de la propuesta que estamos realizando. La extensión propuesta amplía el campo de aplicación de la metodología de esos autores, permitiendo considerar aquellas situaciones donde la forma funcional del efecto de alguna de las variables explicativas sobre la variable de interés es desconocida o situaciones en las que la especificación de una determinada forma funcional resulta un supuesto bastante restrictivo o carece de sentido. Para finalizar, señalaremos que

la inferencia del modelo se ha realizado mediante técnicas bootstrap, para lo cual hemos propuesto un procedimiento de generación de remuestras bootstrap adecuado a las características del modelo.

AGRADECIMIENTOS

Este trabajo ha sido financiado por los proyectos de investigación UPV 038.321-HA129/99 de la Universidad del País Vasco/Euskal Herriko Unibertsitatea, PB98-0149 de la Dirección General de Enseñanza Superior e Investigación Científica del Ministerio Español de Educación y Cultura y PI-1999-70 del Gobierno Vasco/Eusko Jaurlaritza. El autor agradece tanto los comentarios de la editora como de los evaluadores que han servido para mejorar de forma importante el trabajo realizado.

REFERENCIAS

Aitkin, M. & Clayton, D. (1980). «The Fitting of Exponential, Weibull and Extreme Value Distributions to Complex Censored Survival Data using GLIM». *Applied Statistic*, 29, 156-163.

Akritas, M. G. (1986). «Bootstrapping the Kaplan-Meier Estimator». *Journal of the American Statistical Association*, 81, 1032-1038.

Brookmeyer, R. & Gail, M. H. (1993). *AIDS Epidemiology a Quantitative Approach*. Oxford University Press: Oxford.

Buja, A., Hastie, T. J. & Tibshirani, R. J. (1989). «Linear Smoothers and Additive Models (with Discussion)». *Annals of Statistic*, 17, 453-555.

Cox, D. R. (1972). «Regression Models and Life-Tables». *Journal of the Royal Statistical Society-Series B*, 34, 187-220.

Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press: Cambridge.

Efron, B. (1981). «Censored Data and Bootstrap». *Journal of the American Statistical Association*, 76, 312-319.

—— (1987). «Better Bootstrap Confidence Intervals». *Journal of the American Statistical Association*, 82, 171-200.

Efron, B. & Tibshirani, R. (1986). «Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy». *Statistical Science*, 1, 54-77.

—— (1993). *An Introduction to the Bootstrap*. Chapman and Hall: New York.

Good, I. J. & Gaskins, R. A. (1971). «Non-parametric Roughness Penalties for Probability Densities». *Biometrika*, 58, 255-277.

Green, P. J. & Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall: London.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall: London.

Kaplan, E. L. & Meier, P. (1958). «Nonparametric Estimation from Incomplete Observations». *Journal of the American Statistical Association*, 53, 457-481.

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

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*. John Wiley and Sons: New York.

Lemp, G. P., Payne, S. F. & Neal, D. (1990). «Survival Trends for Patients with AIDS». *Journal of the American Medical Association*, 263, 402-406.

McCullagh, P. & Nelder, J. A. (1983). *Generalized Linear Models*. Chapman and Hall: London.

Moore, R. D., Hidalgo, J., Sugland, B. W. & Chaisson, R. E. (1991). «Zidovudine and the Natural History af the Acquired Inmunodeficiency Syndrome». *New England Journal of Medicine*, 263, 1412-1416.

Nelder, J. A. & Wedderburn, R. W. M. (1972). «Generalized Linear Models». *Journal of the Royal Statistical Society-Series A*, 135, 370-384.

Orbe, J. (2000). *Un Modelo de Regresión Parcial Censurado para Análisis de Supervivencia*. Tesis Doctoral, Universidad del País Vasco, Bilbao.

Reid, N. (1981). «Estimating the Median Survival Time». *Biometrika*, 68, 601-608.

Whitehead, J. (1980). «Fitting Cox's Regression Model to Survival Data using GLIM». *Applied Statistics*, 29, 268-275.

ENGLISH SUMMARY

LIFETIME DATA ANALYSIS USING A SEMIPARAMETRIC GENERALIZED LINEAR MODEL

JESUS ORBE*

Aitkin and Clayton (1980) propose to analyze duration models using generalized linear models. In this work, we extend that methodology by allowing the introduction of the effect of some covariable in a nonparametric way. Thus, the proposed model is a semiparametric generalized linear model, with a parametric component where we specify the functional form of the effect of the covariables on the duration variable, and a nonparametric component where we capture the effect of some covariable without assuming any functional form. We develop the estimation process and we use a bootstrap procedure to infer on the estimates for the parameters. As an application, we study the survival time for a sample of AIDS diagnosed patients.

Keywords: Duration models, censorship, bootstrap, generalized linear models, semiparametric estimation

AMS Classification (MSC 2000): 62J12, 62N05, 62G09

* Departamento de Econometría y Estadística (E.A. III). Universidad del País Vasco/Euskal Herriko Unibertsitatea. Avenida Lehendakari Aguirre, 83. 48015 Bilbao. E-mail: jo@alcib.bs.ehu.es.

– Received June 2000.
– Accepted February 2001.

1. INTRODUCTION

In this paper we propose an extension of the work presented by Aitkin and Clayton (1980). These authors put forward the possibility of estimating some duration models using generalized linear models. We use this idea and extend their methodology to a semiparametric case. By using this extension, we can consider situations where we do not know the functional form of the effect of some covariate on the duration or situations where considering some specific parametric functional form for this effect may be very restrictive.

We first describe the link between duration models and generalized linear models. Thus, we consider the class of proportional hazard models

$$\lambda(t;x) = \lambda_0(t)\exp(x^T\beta),$$

where $\eta = x^T\beta$ is the linear predictor. For a sample of n observations, we can obtain this log-likelihood for the model

$$\ln L = \sum_{i=1}^{n} \delta_i \left[\ln \Lambda_0(y_i) + \eta_i\right] - \Lambda_0(y_i)e^{\eta_i} + \delta_i \ln\left(\frac{\lambda_0(y_i)}{\Lambda_0(y_i)}\right)$$

by taking $\mu_i = \Lambda_0(y_i)e^{\eta_i}$, we can rewrite this equation as

$$\ln L = \underbrace{\sum_{i=1}^{n}(\delta_i \ln \mu_i - \mu_i)}_{(a)} + \underbrace{\sum_{i=1}^{n} \delta_i \ln\left(\frac{\lambda_0(y_i)}{\Lambda_0(y_i)}\right)}_{(b)}.$$

It can be verified that the (a) is proportional to the logarithm of the likelihood function of a sample of δ_i random variables with Poisson distribution with mean value μ_i. Therefore, the parameters of the original duration model can be estimated using a generalized linear model (GLM), the log-linear Poisson model,

$$\ln \mu_i = \ln \Lambda_0(y_i) + x_i^T\beta$$

We present a dataset of AIDS diagnosed patients and analyze the effect of some covariables on the survival time from the diagnosis moment using the traditional methodology in survival analysis. Using this application we motivate the extension of the methodology proposed by Aitkin and Clayton (1980).

2. DURATION ANALYSIS USING A SEMIPARAMETRIC GLM

We propose to extend the work of Aitkin and Clayton (1980) by introducing a nonparametric term in the model. Thus, we consider duration models with hazard function

$$\lambda(t;x) = \lambda_0(t)\exp(x^T\beta + h(r)),$$

where $h(r)$ is a smooth function that is used to capture the effect of the covariable R. Therefore, we have passed from a parametric linear predictor $\eta = x^T\beta$ to a semiparametric one $\eta = x^T\beta + h(r)$.

In order to estimate the parameters of the model and the function $h(r)$, we can use the penalized log-likelihood function

$$\Pi = \sum_{i=1}^{n} \left\{ \delta_i \left[\ln \Lambda_0(y_i) + x_i^T\beta + h(r_i) \right] - \Lambda_0(y_i) e^{x_i^T\beta + h(r_i)} + \delta_i \ln \left(\frac{\lambda_0(y_i)}{\Lambda_0(y_i)} \right) \right\} -$$

$$- \frac{1}{2}\alpha \int_a^b \left[h''(r) \right]^2 dr.$$

Here, as in the parametric case, we can use a generalized linear model to obtain the estimators but, in this case, a semiparametric one. Thus, we can use a log-linear Poisson model for the censorship δ indicator variable, with a logarithmic link function and the following semiparametric linear predictor

$$\ln \mu_i = \ln \Lambda_0(y_i) + x_i^T\beta + h(r_i)$$

The penalized log-likelihood function for this «auxiliar» semiparametric generalized linear model is

$$\Pi = \sum_{i=1}^{n} \left[\delta_i \ln \mu_i - \mu_i - \sum_{i=1}^{n} \ln \delta_i! \right] - \frac{1}{2}\alpha \int_a^b \left[h''(r) \right]^2 dr.$$

If we substitute μ_i by $e^{\ln \Lambda_0(y_i) + x_i^T\beta + h(r_i)}$, we can see that this expression is proportional to the previous one.

It can be demonstrated that the solution to maximize the penalized log-likelihood function for $h(r)$ function is a natural cubic spline. Therefore, using some properties of these functions, the penalized log-likelihood can be rewritten as

$$\Pi = \sum_{i=1}^{n} \left\{ \delta_i \left[\ln \Lambda_0(y_i) + x_i^T\beta + (Nh)_i \right] - \Lambda_0(y_i) e^{x_i^T\beta + (Nh)_i} + \delta_i \ln \left(\frac{\lambda_0(y_i)}{\Lambda_0(y_i)} \right) \right\} -$$

$$- \frac{1}{2}\alpha h^T K h$$

The same steps can be carried out in the penalized log-likelihood of the «auxiliar» semiparametric generalized linear model and, then, using the Fisher scoring algorithm, we can obtain the equations system

$$X^T W X \beta = X^T W (Z^* - Nh) \qquad (a)$$

$$(N^T W N + \alpha K) h = N^T W (Z^* - X\beta) \qquad (b)$$

where the i-th element of vector Z^* is

$$z_i^* = \ln \Lambda_0(y_i) + x_i^T \beta + (Nh)_i + (\delta_i - \mu_i)\frac{1}{\mu_i}$$

The elements of the main diagonal in W are

$$w_{ii} = \Lambda_0(y_i) e^{x_i^T \beta + (Nh)_i}$$

In order to obtain the estimators of the model, we can apply a backfitting algorithm between (a) and (b) until convergence is achieved.

3. INFERENCE AND MAIN RESULTS

Once the estimation procedure is finished, we are interested in doing inference. This analysis is done by using bootstrap resampling techniques. In order to do this, we propose a new procedure to obtain the bootstrap resamples that are adequate to the characteristics of our model. Considering our case, a semiparametric exponential regression model, this procedure consists on the following steps:

Step 1: Fit the original model using the proposed methodology.

Step 2: Generate the bootstrap sample for the error variable, $\varepsilon_1^*, \ldots, \varepsilon_n^*$, using a minimum extreme value probability distribution.

Step 3: Obtain the bootstrap sample for the duration variable doing a model-based bootstrap. That is,

$$\ln T_i^* = x_i^T \hat{\beta} + \hat{h}(r_i) + \varepsilon_i^*; \quad \text{for} \quad i = 1, \ldots, n.$$

Step 4: Obtain the bootstrap sample for the censoring variable through the estimation of the distribution function of censoring variable, G.

Step 5: Compare the bootstrap samples of the duration and censoring variables and, thus, obtain the bootstrap sample of the observed variable Y^*, and the corresponding bootstrap indicator variable δ^*,

$$y_i^* = \min\{t_i^*, c_i^*\}, \qquad \text{and} \qquad \delta_i^* = \begin{cases} 1; & \text{if } t_i^* \leq c_i^* \\ 0; & \text{if } t_i^* > c_i^* \end{cases}.$$

Step 6: Estimate the model for the bootstrap sample.

Step 7: Go back to Step 2 and repeat the procedure M times.

After we estimate the model and, using the bootstrap techniques, calculate the standard deviations and confidence intervals, we can summarize the most relevant results obtained from the empirical analysis. With regard to the covariates introduced in the parametric component, the age of the patient has a negative significant effect on the survival time. As for the estimation of the nonparametric component, we can observe that the introduction of AZT treatment has a positive effect on the survival, increasing the survival time of patients. Finally, I would like to add that, with the extension proposed here, it is possible to capture the gradual effect on survival of this medicine, which is not possible by using a dummy variable specification.