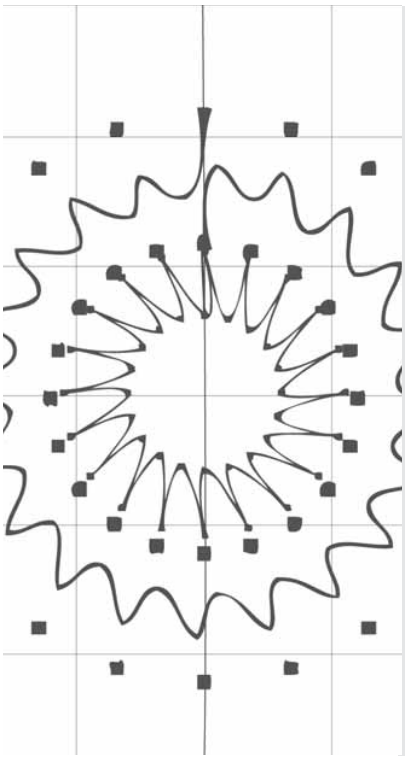
REVISTA Universidad EAFIT
Vol. 46. No. 157. 2010. pp. 11-18

Una Aplicación con datos de sobrevivencia:
Utilización de modelos de regresión parcial censurados



Hermilson Velásquez Ceballos
Doctor en Ciencias Matemáticas,
Universidad Politécnica de Valencia, Valencia, España.
evelas@eafit.edu.co

Héctor Javier Herrera Mejía
Matemático, Universidad de Antioquia.
Magíster en Matemáticas Aplicadas, Universidad EAFIT.
hherrer1@eafit.edu.co

Jorge Iván Jiménez Gómez
Matemático, Universidad de Antioquia.
Magíster en Matemáticas Aplicadas, Universidad EAFIT.
jjimene4@eafit.edu.co

Recepción: 19 de abril de 2009 | Aceptación: 01 de octubre de 2009

* Este artículo es parte del trabajo de Investigación “Análisis Exploratorio de Datos Espaciales y el Índice de Moran”, realizado en 2008 con la financiación de la Universidad EAFIT.

Resumen

Uno de los aspectos centrales en muchas investigaciones en diferentes campos científicos —ingeniería, medicina y economía—, se relaciona con la construcción de modelos que representen de manera adecuada el proceso generador de datos para una variable temporal que recoge información sobre la confiabilidad, sobrevivencia o duración. El trabajo realizado en este aspecto se hizo con base en la metodología que extiende los modelos tradicionales de regresión, utilizados por Cox y Stute en el análisis de sobrevivencia, para incluir en su especificación el efecto de una covariable temporal. Tal análisis se recoge en forma no paramétrica. La parte relevante del estudio lo constituye la forma como se trata el efecto que una variable temporal ejerce sobre la variable asociada con la sobrevivencia. Además, los resultados de Cox, Stute y el modelo de tasa acelerada de falla (AFT) devienen en casos particulares del modelo de regresión parcial censurado.

Palabras clave

Sobrevivencia
Modelo semiparamétrico
Tasa acelerada de falla
Riesgos proporcionales
Curvas *splines*
Modelo de regresión parcial censurado

Use of partial censored regression models: A survival data application

Abstract

A central aspect in research in several scientific fields, such as engineering, medicine and economy, is the construction of models that adequately represent the data generating process for a temporal variable that gathers information regarding reliability, survival or duration. The present work was based in a methodology that extends the traditional regression models used by Cox and Stute in survival analysis, including in their specification the effect of a temporal covariable. The analysis is non-parametrically gathered. The relevant aspect of this work is the form in which the temporal variable effect on the survival variable is treated. Furthermore, the results from Cox and Stute and the Accelerated Failure Time (AFT) model become particular cases of the partial censored regression model.

Key words

Survival
Semi-Parametric Model
Accelerated Failure Time
Proportional Risk
Spline Curves
Partial Censored Regression Model

Introducción

En diferentes campos científicos como economía, medicina, ingeniería y biología, uno de los objetos de estudio está asociado con la modelación de alguna variable que hace referencia a la duración, sobrevivencia o confiabilidad. A través de la historia se han realizado múltiples estudios relacionados con la sobrevivencia de personas, animales, duración de una máquina, de una huelga o de un componente, etc. Todos esos estudios tienen en común que la investigación está centrada en el análisis de la variable duración o sobrevivencia. Ejemplos de este tipo de fenómenos son el tiempo de sobrevivencia de un individuo desde que se le diagnóstica una enfermedad hasta que muere, tiempo de duración de un componente de un dispositivo electrónico hasta que falla, duración de una huelga, etc.

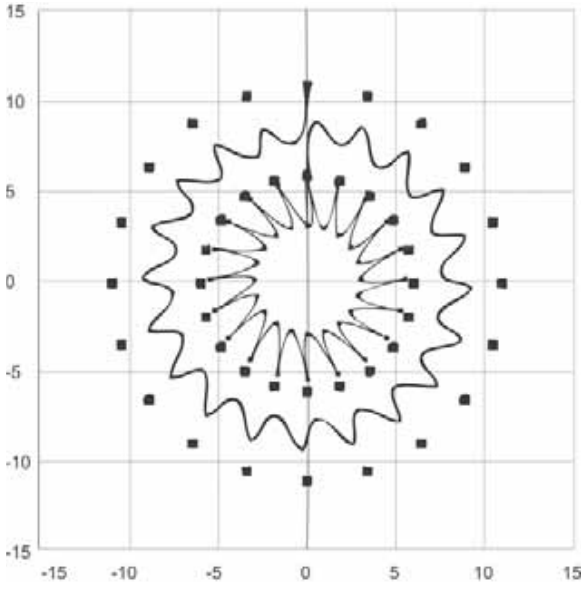
Un aspecto fundamental a tener en cuenta en estos procesos está relacionado con el tipo de datos utilizados, en los cuales, por su misma naturaleza, está presente la censura, ya sea por la muerte, falla, desaparición de los individuos u objetos de estudio.

El tiempo de duración o de sobrevivencia tiene la misma connotación en ingeniería, donde se maneja con el nombre de confiabilidad; en medicina se le conoce como sobrevivencia, y en economía se denomina duración.

Una primera aproximación a la modelación de este tipo de procesos está relacionada con el ajuste de una distribución teórica de los datos. Entre las diferentes distribuciones teóricas se encuentran la binomial, gama, *poisson*, *weibull* y exponencial. El análisis que se hace de esta información se enriquece sustancialmente si se complementa con otros procedimientos que recurren a nuevos enfoques para modelar este tipo de fenómenos.

En tal sentido, es natural pensar que el tiempo de duración puede depender de otras variables que ayuden a su explicación; en consecuencia, Cox (1972) utilizó un modelo que llamó de riesgos proporcionales (PH) para modelar su proceso generador. En dicho modelo se tiene una limitación: debe cumplir con los supuestos de riesgos proporcionales, con las consecuencias que esto conlleva, aunque tiene la ventaja de que no se necesita conocer la distribución de los datos.

El modelo de tasa acelerada de falla (AFT), propuesto por Stute (1999), no necesita asumir riesgos proporcionales ni una distribución para el tiempo de sobrevivencia, haciéndolo más amigable y con buenas propiedades estadísticas.



Este nuevo enfoque aparece en los trabajos pioneros sobre regresión parcial censurada, realizados por el doctor Vicente Núñez Antón (Núñez-Antón & Orbe, 2004; Orbe, Ferreira & Núñez-Antón, 2002 y 2003; Orbe & Núñez-Antón, 2006), en los cuales los modelos de Cox y Stute resultan ser casos particulares ya que no requieren los supuestos de riesgos proporcionales o el conocimiento de la distribución para la variable duración y, además, de su especificación se puede conseguir la representación funcional de dichos modelos.

Bajo tal perspectiva, el objetivo central de este artículo está relacionado con la presentación y aplicación de los resultados más recientes que se han obtenido para modelar este tipo de fenómenos, entre los cuales un aspecto fundamental lo constituye el hecho de recoger el efecto que sobre la variable de sobrevivencia ejerce una covariable cuya información es temporal.

El artículo se compone de seis apartados, además de esta introducción. Una síntesis del modelo de regresión parcial censurado, sección donde se abordan desde el punto de vista teórico y matemático los modelos ya conocidos y se hace un resumen de los aspectos más importantes de los mismos y una breve mención de los modelos de regresión parciales censurados y las referencias utilizadas para su construcción. El siguiente hace referencia a la estimación del modelo propuesto. Posteriormente, se presenta el caso de aplicación y, enseguida, se hace la descripción del problema, para luego pasar al análisis de los resultados obtenidos. Por último, se recogen las conclusiones de la investigación, a la vez que se muestran las puertas para futuras investigaciones que, a criterio de los autores, quedan abiertas después de la realización de este estudio.

Modelo de regresión parcial censurado

Los modelos que se agrupan bajo el concepto de regresión parcial censurado extienden y generalizan las metodologías desarrolladas por Cox (1972) y Stute (1999). Uno de los aspectos centrales y más importantes de este tipo de modelos es que permiten capturar la dinámica que sobre la variable duración ejerce una covariable cuya información es una serie de tiempo.

La especificación relacionada con este tipo de procesos es:

$$Y_i = X_i\beta + h(Z_i) + \varepsilon \quad [1]$$

Donde:

X_{nxp} : Matriz con las observaciones de las variables que se consideran paramétricas en la especificación

h : Es una curva de suavización *spline* cúbica

Z_{nx1} : Es el vector con la información de la variable no paramétrica

β_{px1} : Vector de parámetros

Estimación de los parámetros en el modelo propuesto

A partir de la especificación [1] se resuelve el problema de optimización que permita encontrar estimaciones para β y *h*.

$$\min_{h,\beta} \sum_{i=1}^{n} \varepsilon_i^2 = \min_{h,\beta} \sum_{i=1}^{n} [Y_i - X_i\beta - h(Z_i)]^2 \quad [2]$$

El procedimiento anterior no es directo debido a la forma como se considera *h*.

El procedimiento de solución considera inicialmente la suma de cuadrados ponderada:

$$S_W(\beta,h) = \sum_{i=1}^{n} W_{in}[Y_i - X_i\beta - h(Z_i)]^2 + \alpha \int [h''(z)]^2 dz \quad [3]$$

La expresión:

$$\sum_{i=1}^{n} W_{in}[Y_i - X_i\beta - h(Z_i)]^2 \quad [4]$$

Se puede escribir en forma ordenada así:

$$\sum_{i=1}^{n} W_{in}[Y_{[i]} - X_{[i]}\beta - h(Z_{[i]})]^2 \quad [5]$$

Y en forma matricial como:

$$\sum_{i=1}^{n} W_{in}[Y_{[i]} - X_{[i]}\beta - h(Z_{[i]})]^2 = (Y - X\beta - Nh)^T W (Y - X\beta - Nh) \quad [6]$$

El vector *h* contiene los valores $h(s_j)$ para los distintos valores ordenados $s_1 < s_2 < s_3 < \ldots < s_d$, de la covariable *Z*.

La conexión entre $s_1 < s_2 < s_3 < \ldots < s_d$ y $z_1 < z_2 < < z_3 < \ldots < z_n$ se da a través de las componentes de la matriz *N* de orden *nxd,* la cual asigna los valores respectivos de la covariable *Z* para el *i*-ésimo individuo, es decir,

$$N_{ij} = \begin{cases} 1 & si \;\; z_i = s_j \\ 0 & en\ otro\ caso \end{cases}$$

Para la suavidad se recurre a la integral para el cuadrado de la segunda derivada, entonces la combinación de ambos términos puede ser utilizada para minimizar las ponderaciones de la expresión de

mínimos cuadrados que, escrita en forma matricial, sería así:

$$S_W(\beta,h)=(Y-X\beta-Nh)^T W(Y-X\beta-Nh)+\alpha\int[h''(z)]^2dz \qquad [7]$$

Para minimizar $S_w(\beta,h)$ se pueden considerar dos pasos: primero, se minimiza $h(s_j)$ para j=1,2,3,...,d; segundo, se minimiza el resultado, a partir de iterar varias veces entre β y h.

El problema de minimizar $\alpha\int[h''(z)]^2dz$ está sujeto a la interpolación, dados los puntos $h(s_j)$, lo cual produce una *spline* cúbica con nodos $\{s_j\}$.

El grado de suavidad está determinado por α, el parámetro suavizador. Grandes valores de α producen curvas suaves, mientras pequeños valores producen curvas más inestables. Dado $\alpha > 0$, la solución que minimiza la ecuación anterior es una *spline* cúbica. Una función con nodos en $s_1 < s_2 < s_3 < ... < s_d$. Estos son consecuencia matemática de la selección de $\alpha\int[h''(z)]^2dz$ que es el término penalizador.

Cuando se aproxima a cero, el término penalizador pierde importancia y la solución tiende a la *spline* cúbica que interpola los datos, es decir, es tal que $h(Z_{[i]})=Y_{[i]}-X_{[i]}\beta$ o el promedio, si están enlazadas. Sin embargo, cuando α es lo bastante grande, el término penalizador domina y así se obtiene la solución de los mínimos cuadrados ponderados para el modelo lineal, es decir, la solución provee una función lineal para $h(\cdot)$. En la práctica, los valores de α pueden ser seleccio-nados usando el criterio de los *data-driven*.

Si se va a utilizar un método empírico, se maneja el método de validación cruzada y este sería el criterio para el análisis.

Si se utilizan las propiedades de las *splines* cúbicas, la integral se puede calcular como $\alpha h^T kh$; donde k es una matriz cuadrada de orden d que se calcula como lo hacen Green y Silverman (1994); solo depende de la localización de los nodos. La matriz h se encuentra como se calculó antes. Así, la minimización del problema se puede escribir de esta forma:

$$S_W(\beta,h)=(\mathbf{Y}-X\beta-Nh)^T W(\mathbf{Y}-X\beta-Nh)+\alpha h^T kh \qquad [8]$$

Por simple inspección o completando cuadrados, la ecuación [8] se minimiza cuando β y h satisfacen la ecuación matricial:

$$\begin{bmatrix} X^TWX & X^TWN \\ N^TWX & N^TWX+\alpha K\end{bmatrix}\begin{pmatrix}\beta\\ h\end{pmatrix}=\begin{bmatrix}X^T\\ N^T\end{bmatrix}W\mathbf{Y} \qquad [9]$$

[9] forma un sistema de $p+d$ ecuaciones que se puede reescribir como un sistema de ecuaciones simultáneas:

a) $X^TWX\beta = X^TW(\mathbf{Y} - Nh)$

b) $(NhW + \alpha k)h = N^TW(\mathbf{Y} - X\beta)$ [10]

Donde:

X : Matriz de información paramétrica

Y : Matriz de sobrevivencia

N : Matriz de nodos coincidentes

W: Matriz de ponderaciones Kaplan Meier

Los detalles formales de estos procesos aparecen en la tesis de maestría de Herrera Mejía y Jiménez Gómez (2008).

En el sistema de ecuaciones [10], la ecuación **a**) dice que si h es conocida, se puede encontrar $(Nh)_i = h(s_i)$ para Y_i y estimar β por una ponderación de la regresión de los mínimos cuadrados de las diferencias. Similarmente, si β es conocido, la ecuación **b**) se ajusta a la suavización de una *spline* cúbica para las diferencias $Y_i - X_i^T\beta$.

La unicidad en la solución se da porque

1) W es diagonal, con elementos no negativos y además es una matriz simétrica
2) Las columnas de X son linealmente independientes
3) No hay combinaciones lineales de $X\beta$ igual a una forma lineal $\delta_2 + \delta_2 Z$.

En la práctica, se pueden obtener estimadores para β y h iterando las dos ecuaciones anteriores hasta alcanzar una convergencia deseada; esto se puede lograr, por ejemplo, haciendo uso del algoritmo de *backfitting*.

Las condiciones 1 y 3 también garantizan la convergencia del algoritmo. Esta convergencia es bastante rápida en muchas situaciones.

Caso de aplicación

Debido a la poca disponibilidad de este tipo de información en Colombia, los datos empleados para el estudio se tomaron de una base registrada en University of Massachusetts[1].

El tamaño de la muestra fue de 1.000 observaciones con nueve variables, de las cuales se utilizaron siete. Su definición aparece a continuación:

id_i : Identificador para el paciente i. i = 1,…., 1.000.

Y_i : Número de días de seguimiento para el paciente i. i = 0.5,…., 180.

X_{1i} : Número de días que el paciente *i* es sometido a la revascularización.

X_{2i} : Años que tiene el paciente al momento de ingresar al hospital entre (28 y 96 años).

X_{3i} : Variable dicótoma que toma los valores 1, si el paciente i presenta desviación con respecto al segmento ST en el electrocardiograma; 0, si no presenta tal desviación.

X_{4i} : Variable dicótoma. 0, si el paciente i muere durante el seguimiento; 1, si el paciente no muere.

X_{5i} : Número de días de permanencia del paciente *i* en el hospital: *i* = 0,.., 52.

En concordancia con la teoría presentada se tiene:

Y: Variable asociada a la sobrevivencia.

X_1, X_2, X_3 : Estas son las variables cuyo efecto se recogerá en forma paramétrica.

X_5 : La variable cuyo efecto se recogerá en forma no paramétrica. Esta variable está relacionada con el eje central del artículo; con ella se captura la dinámica que ejerce X_5 sobre la variable de sobrevivencia *Y*.

La especificación general del modelo [1] es:

$$Y_i = X_i\beta + h\,(Z_i) + \varepsilon i$$

En este caso particular, la especificación propuesta es:

$$Ln(Y) = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \qquad [11]$$

$\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ son las estimaciones de los parámetros asociados a las variables que el modelo recoge en forma paramétrica.

La estimación de este tipo de modelos presenta resultados asociados con la componente paramétrica y también con la componente no paramétrica.

1. En relación con la parte paramétrica, las estimaciones obtenidas en su respectivo orden fueron:

Betas para alfa = 5

$$\hat{\beta}_1 = 0.0220,\ \hat{\beta}_2 = 0.0253,\ \hat{\beta}_3 = -0.0995$$

Betas para alfa = 10

$$\hat{\beta}_1 = 0.0220,\ \hat{\beta}_2 = 0.0253,\ \hat{\beta}_3 = -0.0995$$

La parte paramétrica presenta valores que son robustos para el parámetro de suavización determinado por alfa. Esto se puede observar ya que la estimación de los betas no cambia para valores diferentes de alfa, es decir, los valores de los betas son estables (se manifiestan solo para dos valores de alfa). Se puede notar también que un incremento de una unidad en el tiempo de revascularización incrementa en 0.022% días el tiempo de sobrevivencia, lo cual indica que la revascularización afecta de manera positiva la sobrevivencia.

Estos resultados también muestran que la edad afecta positivamente la sobrevivencia, hecho que solo es cierto de manera relativa, ya que después de determinada edad es de esperar que esta variable influya en forma negativa para la sobrevivencia.

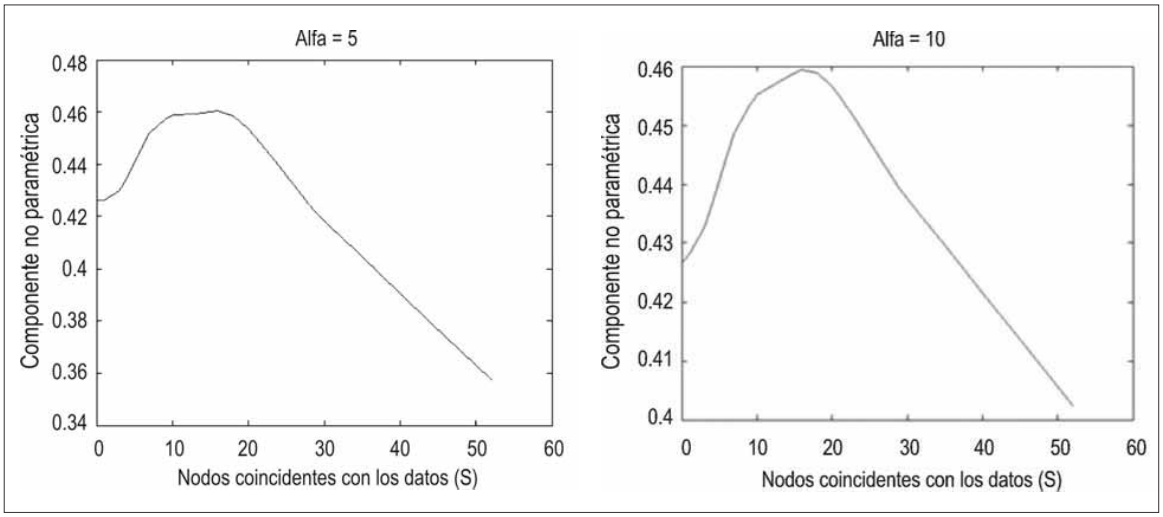
1 La base de datos se puede consultar en www.outcomes-umassmed.org/grace. También se encuentra disponible en ftp// ftp.wiley.com/public/sci_tech_med/survival Global Registry of Acute Coronary Events (GRACE). Fue facilitada por el doctor Frederick Anderson, Jr., director del Center for Outcomes Research (COR), University of Massachusetts, Worcester. Nombre: Global Registry of Acute Coronary Events Data (grace1000.dat)

También se puede observar que para un incremento de una desviación entre la onda *S* y la onda *T* en el electrocardiograma, existe una disminución de 0.0995% en los días de sobrevivencia, lo cual indica que las desviaciones en este segmento afectan de manera negativa la sobrevivencia.

2. En cuanto a la componente no paramétrica, cuya información es capturada por $h(z)$ y que muestra la dinámica de X_5 (número de días de permanencia en el hospital) sobre *Y,* su comportamiento se recoge en forma gráfica en la figura 1.

Las dos gráficas permiten observar un patrón creciente en el tiempo de sobrevivencia cuando el tiempo máximo de permanencia en el hospital es de 20 días. A partir de este valor, la permanencia no incrementa la sobrevivencia y el fenómeno analizado tiende a seguir su proceso natural. Que la influencia de X_5 sobre la sobrevivencia se vea reflejada en una curva y no en un valor numérico constituye un aporte relevante para la modelación de procesos de duración. Este es un aspecto fundamental que permite enriquecer y extender los análisis tradicionales que usualmente se realizan con este tipo de información.

Figura 1. Comportamiento de la componente no paramétrica



Conclusiones

El caso de estudio que se ha presentado extiende el análisis de sobrevivencia tradicional. En este, solo se consideran ajustes de distribuciones de probabilidad o análisis de regresión, en los cuales se recogen los efectos de las covariables en forma paramétrica a una nueva dimensión. Aquí, además, se incluyen, de manera específica, las nuevas variables que aportan información relevante para el análisis de sobrevivencia. Lo anterior permite mostrar el efecto que, en general, las variables relacionadas con el tiempo pueden ejercer sobre la variable de sobrevivencia, tal como se aprecia en el comportamiento de la variable X_5.

La parte relevante y diferente de este trabajo lo constituye la forma como fue recogido el efecto que la variable X_5 ejerce sobre la sobrevivencia.

Los resultados tradicionales —Cox, Stute y el AFT—, aparecen como casos particulares en esta especificación.

El tratamiento computacional asociado con el modelo propuesto no requiere de *software* especializado para el procesamiento de la información; las estimaciones que se presentaron fueron implementadas en *Matlab*.

Ahora bien, con ese mismo propósito se puede explorar la utilización de suavizadores del tipo Kernel, modelos Arima y otros tipos de *Splines* en trabajos futuros con los cuales se pretenda recoger en forma no paramétrica la información de la variable temporal.

Bibliografía

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

_____. (1972). "Regression models and life-tables", *Journal of the Royal Statistical Society, 34*(Series B).Publicado por: Wiley-Blackwell, pp. 187-220.

Green, P. J & B. W. Silverman. (1994). *Nonparametric regression and generalized linear models.* London: Chapman and Hall.

Herrera Mejía, Héctor y Jorge Jiménez Gómez. (2008). "Modelo de regresión semiparamétrico con datos censurados: fundamentos y una aplicación". Tesis de grado para optar el título de Magíster en Matemáticas Aplicadas. Universidad EAFIT, Medellín.

Kaplan, E. l. & P. Meier. (1958). "Nonparametric estimation from incomplete observations", *Journal of the American Statistical and probability, 53.* Publicado por: American Statistical Association, pp. 457-481.

Núñez-Antón, Vicente & Jesús Orbe. (2005). "Statistical time to event analysis in the social sciences", *Modeling Hazard Rate and duration in Finance*. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences Volumen 1, Número 3, pp.103-117

Orbe, Jesús. (2000). "Un modelo de regresión parcial censurado para análisis de supervivencia". Tesis doctoral, Universidad de País Vasco, Bilbao, España.

Orbe, Jesús & Vicente Núñez-Antón. (2006). "Alternative approaches to Study life time data under different scenarios: from the PH to the modified semiparametric AFT model", *Computational Statistics & Data Analysis, 50*. Elsevier Science Publishers B. V., pp. 1565-1582.

Orbe, Jesús; Ferreira, Eva & Núñez-Antón Vicente. (2003). "Censored partial regression", *Biostatistics, 41*. Oxford University Press, pp. 109-121.

_____. (2002). "Comparing proportional hazards and accelerated failure time model for survival analysis", *Statistics in Medicine, 21*., pp. 3493-3510.

Stute, W. (1999). "Nonlinear censored regression", *Statistica Sinica, 9.*, pp. 1089-1102. Disponible en [línea]: http://www3.stat.sinica.edu.tw/statistica/oldpdf/A9n413.pdf Consultado en: 10/10/2009.

_____. (1996a). "Distributional convergence under random censorship when covariables are present", *Scandinavian Journal of Statistics, 23*. Publicado por: Blackwell Publishing, pp. 461-471.

_____. (1996b). "The jack-knife estimate of the variance of the Kaplan-Meier integral", *Annals of Statistics, 24*. IMS Journals and Publications, pp. 2679-2704.

_____. (1993). "Consistent estimation under random censorship when covariables are present", *Journal of Multivariate Analysis, 45*. Elsevier Science Publishers B. V., pp. 89-103.