ESTIMACIÓN DE LA SUPERVIVENCIA CONDICIONAL BAJO SESGO LONGITUDINAL Y CENSURA TIPO I

Jacobo de Uña Álvarez¹ y M^a del Carmen Iglesias Pérez¹

¹Departamento de Estadística e Investigación Operativa
Universidad de Vigo

RESUMEN

De Uña-Álvarez (2003) propuso un estimador de la función de supervivencia bajo sesgo longitudinal y censura tipo I. En este trabajo se adaptan estas ideas a la estimación de curvas de supervivencia condicionales. El trabajo incluye el estudio asintótico de un nuevo estimador de la supervivencia condicional, así como una ilustración con datos reales.

Palabras clave: sesgo longitudinal, censura, supervivencia condicional, estimación no paramétrica, e...ciencia.

1. INTRODUCCIÓN

El análisis de datos de duración tiene interés en distintos campos aplicados, como la biomedicina, la ingeniería, o la socioeconomía. En estudios de este tipo, la recogida de datos se encuentra condicionada por problemas en el muestreo, tales como los ocasionados por los fenómenos de censura y truncamiento. En este trabajo nos centramos en muestreos por corte transversal; es decir, los períodos de duración observados son aquellos que se encuentran "en marcha" en un instante temporal, en el cual el observador llega al proceso. Este muestreo por corte transversal provoca un truncamiento por la izquierda. Si denotamos la variable T como el tiempo transcurrido desde el evento inicial (diagnóstico, por ejemplo) hasta la llegada del observador, entonces se observa la variable de interés Y si y sólo si $T \leq Y$. Cuando existe un límite temporal en el seguimiento del sujeto, se produce un riesgo de censura por la derecha. Pongamos $C = T + \zeta$, donde ζ es una constante positiva (conocida) que representa la duración del seguimiento. Entonces, se observa $(T; Z; \pm)$ si y sólo si $T \leq Z$, donde $Z = \min(Y; C)$ y $\pm = 1_{\{T \leq C\}}$. Ésta es la situación de censura Tipo I considerada por Winter y Földes (1988) y Wang (1991), entre otros.

Denotemos por F y L las funciones de distribución de las variables Y y T , respectivamente. Cuando las variables T e Y son independientes, el estimador no paramétrico de máxima verosimilitud de F es el estimador de Turnbull (véase Tsai et al. (1987)). Sin embargo, cuando se tiene información auxiliar sobre L , se pueden construir estimadores más eficientes. En particular, esta situación es realista en el caso estacionario, en el cual la incidencia de los eventos iniciales es constante a lo largo del tiempo. El interés de estas aplicaciones ha sido destacado por distintos autores (Wang (1991), Asgharian et al. (2002)). La estacionariedad del proceso subyacente indica que la variable T se encuentra uniformemente distribuida, situación que lleva a un modelo de sesgo longitudinal en la variable Y de interés (Vardi (1982)):

$$F^{\alpha}(y) = P(Y \leq y | T \leq Z) = \int_0^Z \frac{1}{F} dF(u); \quad (1)$$

donde $\bar{F}$ denota la media de F . La ecuación (1) permite introducir un estimador de F a partir de un estimador de su distribución sesgada por longitud F^* (Asgharian et al. (2002), de Uña-Álvarez (2003)).

En la situación descrita, es fácil verificar que

$$F(y) = \bar{F} \frac{\int_0^y u^{i-1} dF^*(u)}{\int_0^y u^{i-1} dF^*(u)}$$

donde además

$$F^*(y) = E \left[\frac{1_{f_{Z \cdot y} \leq g}}{p(Z)} \mid T \cdot Z \right], \quad \text{con } p(y) = 1_{f_{y < \zeta g}} + \frac{\zeta}{y} 1_{f_{y \leq \zeta g}} \quad (2)$$

Entonces, dados $(T_1; Z_1; \pm_1); \dots; (T_n; Z_n; \pm_n)$ independientes e idénticamente distribuidos, con la misma distribución que $(T; Z; \pm)$ condicionada a $T \cdot Z$, un estimador natural de F está dado por

$$\hat{F}(y) = \frac{\int_0^y u^{i-1} d\hat{F}^*(u)}{\int_0^y u^{i-1} d\hat{F}^*(u)} \quad \text{donde } \hat{F}^*(y) = \frac{1}{n} \sum_{i=1}^n \frac{1_{f_{Z_i \cdot y} \leq \pm_i}}{p(Z_i)} \quad (3)$$

(de Uña-Álvarez (2003)). El estimador $\hat{F}$ es una generalización al contexto de censura del empírico en Vardi (1982), que resulta ser más eficiente que el estimador de Turnbull. En comparación con otras generalizaciones (Asgharian et al. (2002)), goza de forma explícita simple, lo cual facilita su implementación así como el estudio de las propiedades asintóticas.

Un problema crucial es el de la extensión del estimador (3) en presencia de covariables. En este caso, se asume que el vector observable es de la forma $(X; T; Z; \pm)$, donde X es un vector de variables explicativas, y que aquél se observa cuando y sólo cuando $T \cdot Z$. En este trabajo consideramos el problema de la estimación no paramétrica de la función de distribución condicional $F(y \mid x) = P(Y \leq y \mid X = x)$ bajo la situación de sesgo longitudinal y censura Tipo I descrita arriba. Para ello, se propone una adecuada generalización del modelo (Modelo 1) y el estimador en de Uña-Álvarez (2003) para albergar el vector de covariables. Como resultados principales, se prueba la consistencia uniforme del estimador introducido (incluyendo órdenes de convergencia), así como su normalidad asintótica. Asimismo, se demuestra que el nuevo estimador es (asintóticamente) más eficiente que el estimador condicional en Iglesias-Pérez y González-Manteiga (1999), el cual no hace uso de la información adicional sobre la distribución de truncamiento. Por último, se ilustra el nuevo estimador utilizando datos reales.

2. DEFINICIÓN DE UN ESTIMADOR DE LA SUPERVIVENCIA CONDICIONAL

En esta sección se revisa el modelo de censura y sesgo longitudinal en de Uña-Álvarez (2003), con el objetivo de definir un estimador de la función de distribución (y supervivencia) condicional. El modelo de sesgo longitudinal y censura tipo I con covariables (Modelo 2) asume las siguientes hipótesis:

- i) Y y T son condicionalmente independientes a X ;
 - ii) $T \mid X = x$ se distribuye uniformemente en $(0; \zeta L_x)$
 - iii) $C = T + \zeta$;
 - iv) $0 < a_{F_x}$ y $b_{F_x} < \zeta L_x$ (siendo a_{F_x} y b_{F_x} los extremos del soporte de $F(y \mid x)$) y además existe $\int_{a_{F_x}}^{b_{F_x}} s dF(s \mid x) > 0$;
- (Estas hipótesis se cumplan para ciertos valores x que más adelante precisaremos).

Bajo las hipótesis i)-iv) se verifica que:

$$H_1^n(y|x) = \frac{P(Z \cdot y; \pm = 1 | Z \leq T; X = x)}{P(T \cdot Y | X = x)} = \frac{P(Y \cdot y; Y_i \leq T \cdot Y | X = x)}{P(T \cdot Y | X = x)}$$

$$= \frac{\int_0^y \frac{1}{L_x} 1_{f_{S_i} \leq 0g} + \frac{s}{L_x} 1_{f_{S_i} < 0g} dF(s|x)}{\int_0^1 \frac{s}{L_x} dF(s|x)}$$

Y por tanto:

$$dH_1^n(s|x) = \frac{w(s)}{\int_0^1 s dF(s|x)} dF(s|x); \quad (4)$$

donde $w(s) = s 1_{f_{S_i} < 0g} + \int 1_{f_{S_i} \leq 0g}$. (Nótese que $w(y) = yp(y)$ siendo $p(y)$ la función de densidad en (2). Nótese también que (4) implica $a_{F_x} = a_{H_1^n}$).

La relación (4) permite escribir $F(y|x)$ como función de cantidades estimables, en la forma siguiente:

$$F(y|x) = \int_0^y dF(s|x) = \int_0^y s dF(s|x) \int_0^y \frac{dH_1^n(s|x)}{w(s)}; \quad (5)$$

donde además:

$$1_{F_x} = \int_0^1 s dF(s|x) = \int_0^1 \frac{dH_1^n(s|x)}{w(s)}; \quad (6)$$

Si la covariable X es continua, un estimador razonable de $H_1^n(y|x)$ es el estimador no paramétrico con pesos de Nadaraya-Watson, dado por:

$$\hat{H}_1^n(y|x) = \frac{1}{n} \sum_{i=1}^n 1_{f_{Z_i} \cdot y; \pm_i = 1g} B_{ni}(x); \quad \text{donde} \quad B_{ni}(x) = \frac{K \left(\frac{x_i - x}{h} \right)}{\sum_{j=1}^n K \left(\frac{x_j - x}{h} \right)}$$

siendo K una función núcleo y h el parámetro ventana.

Entonces, teniendo en cuenta (5) y (6), y sustituyendo $H_1^n(y|x)$ por $\hat{H}_1^n(y|x)$ en dichas expresiones, podemos definir un estimador de $F(y|x)$ que resulta ser:

$$\hat{F}_h(y|x) = \int_0^y \frac{d\hat{H}_1^n(s|x)}{w(s)} \int_0^y \frac{d\hat{H}_1^n(s|x)}{w(s)} = \frac{\sum_{i=1}^n 1_{f_{Z_i} \cdot yg} \frac{B_{ni}(x)}{w(Z_i)}}{\sum_{j=1}^n \frac{B_{nj}(x)}{w(Z_j)}} \quad (7)$$

donde $w(Z_j) = Z_j$, si $Z_j < \int$ y $w(Z_j) = \int$, si $Z_j \geq \int$.

A partir del estimador $\hat{F}_h(y|x)$ se definen, de forma inmediata, (entre otros) estimadores de:

1. La función de supervivencia condicional $S(y|x) = 1 - F(y|x)$:

$$\hat{S}_h(y|x) = 1 - \hat{F}_h(y|x);$$

2. La función de regresión $f(y|x) = E(Y | X = x)$:

$$\hat{f}_h(y|x) = \int y d\hat{F}_h(y|x) = \frac{\sum_{i=1}^n Z_i \frac{B_{ni}(x)}{w(Z_i)}}{\sum_{j=1}^n \frac{B_{nj}(x)}{w(Z_j)}};$$

3. La mediana condicional $Me(y|x) = \inf \{y : F(y|x) \geq 0.5\}$:

$$\hat{M}_{eh}(y|x) = \inf \{y : \hat{F}_h(y|x) \geq 0.5\};$$

3. PROPIEDADES DEL ESTIMADOR

En la demostración de los resultados de esta sección, se utilizan las siguientes hipótesis:
 (H1) X es una variable aleatoria real y absolutamente continua.

Sean $M(x) = P(X \leq x)$ y $M^m(x) = P(X \leq x | T \leq Z)$. Denotaremos por m y m^m a las densidades asociadas a M y M^m , respectivamente.

(H2)a) Sea $I = [x_1; x_2]$ un intervalo contenido en el soporte de m^m , verificando que existe un entorno $I_{\pm} = [x_1 \pm \epsilon; x_2 \pm \epsilon]$ con $\epsilon > 0$ tal que:

$$0 < \epsilon = \inf [m^m(x) : x \in I_{\pm}] < \sup [m^m(x) : x \in I_{\pm}] = \epsilon < 1$$

y $0 < \epsilon_j < 1$.

b) Además, para todo $x \in I_{\pm}$; se verifica:

1) Las hipótesis i)-iv) del Modelo 2.

2) Existe $a > 0$ tal que $a \leq a_{F_x}$:

(H3) Las funciones $m(x)$ y 1_{F_x} tienen primera derivada respecto a x continua en $I_{\pm}$ y $F(y|x)$ tiene primera derivada respecto a x continua y acotada en $(y|x) \in [0; 1] \cap I_{\pm}$:

(H4) Las funciones $m(x)$ y 1_{F_x} tienen segunda derivada respecto a x continua en $I_{\pm}$ y $F(y|x)$ tiene segunda derivada respecto a x continua y acotada en $(y|x) \in [0; 1] \cap I_{\pm}$:

(H5) La función núcleo K , es una función de densidad simétrica, que vale 0 fuera del intervalo $(-1; 1)$ y su variación total está acotada por $\int_{-1}^1 K(z) dz < +1$:

(H6) El parámetro ventana $h = (h_n)$ verifica: $h \rightarrow 0$; $\ln n = o(nh)$ y $nh^5 = o(\ln n) = o(1)$:

Teorema 1 (Orden de convergencia uniforme fuerte) Asumiendo un Modelo 2 y bajo las hipótesis H1-H6 se verifica:

$$\sup_{y, a; x \in I} |F_h(y|x) - F(y|x)| = O\left(\frac{\tilde{A} \mu \ln n}{nh}\right) \quad a.s.$$

Demostración. Se obtiene a partir de la descomposición:

$$\begin{aligned} F_h(y|x) - F(y|x) &= \int_0^{\mu Z} \frac{dH_1^m(s|x)}{w(s)} \mathbb{1}_{i=1} - \int_0^Z \frac{dH_1^m(s|x)}{w(s)} \mathbb{1}_{i=1} + \int_0^Z \frac{dH_1^m(s|x)}{w(s)} \mathbb{1}_{i=2} \\ &+ \int_0^Z \frac{dH_1^m(s|x)}{w(s)} \mathbb{1}_{i=3} - \int_0^{\tilde{A}Z} \frac{dH_1^m(s|x)}{w(s)} \mathbb{1}_{i=1} + \int_0^{\mu Z} \frac{dH_1^m(s|x)}{w(s)} \mathbb{1}_{i=1} \mathbb{1}_{i=5}; \end{aligned}$$

sin más que tener en cuenta: la existencia de 1_{F_x} , la acotación $a < w(y) < \epsilon$ para $y \in a$, y

$$\sup_{y, 0; x \in I} |H_1^m(y|x) - H_1^m(y|x)| = O\left(\frac{\tilde{A} \mu \ln n}{nh}\right) \quad c.s. \quad (8)$$

((8) es consecuencia del Lemma 5 en Iglesias-Pérez y González-Manteiga (1999)).

Teorema 2 (Normalidad asintótica) Asumiendo un Modelo 2 y bajo H1-H6 se verifica:

$$\sqrt{nh} (F_h(y|x) - F(y|x)) \xrightarrow{d} N(b(y|x); \frac{1}{2} b(y|x)) \quad \text{para } x \in I \text{ e } y \in a;$$

donde:

$$b(y|x) = \int_{-1}^1 \frac{1_{F_x}}{2m^m(x)} \int z^2 K(z) dz (2^{(0)}(x)m^{(0)}(x) + \mathbb{1}^{(0)}(x)m^m(x));$$

siendo $\mathcal{C}(u) = \int_{F_u}^1 [F(y|j|u) - F(y|j|x)]$; y

$$\mathcal{H}^2(y|x) = \frac{1}{m^n(x)} \int_{\mu_Z} \int_{\mathbb{R}^1} K^2(z) dz \int_0^1 (1 - 2F(y|j|x)) \int_0^Z \frac{dF(s|x)}{w(s)} + F^2(y|x) \int_0^Z \frac{dF(s|x)}{w(s)};$$

Demostración. La siguiente descomposición:

$$\begin{aligned} \hat{F}_{h|j} - F &= \int_0^{\mu_Z} \int_{\mathbb{R}^1} \frac{dH_1^n}{w} \int_0^Z \frac{dH_1^n}{w} - \int_0^{\mu_Z} \int_{\mathbb{R}^1} \frac{dH_1^n}{w} \int_0^Z \frac{dH_1^n}{w} + \\ &+ \int_0^{\mu_Z} \int_{\mathbb{R}^1} \frac{dH_1^n}{w} \int_0^Z \frac{dH_1^n(s|x)}{w(s)} - \int_0^{\mu_Z} \int_{\mathbb{R}^1} \frac{dH_1^n}{w} \int_0^Z \frac{dH_1^n}{w} \\ &= S_1 + S_2; \end{aligned}$$

y el resultado $\rho_{nh} S_2 = o_p(1)$ (consecuencia de (8)) implican la igualdad de las distribuciones asintóticas de las variables $\rho_{nh} \hat{F}_{h|j} - F$ y $\rho_{nh} S_1$.

El estudio de la distribución límite de $\rho_{nh} S_1$ se realiza siguiendo el mismo esquema que el utilizado en la prueba del Corolario 3 en Iglesias-Pérez y González-Manteiga (1999). Para ello resulta de interés notar que:

$$\rho_{nh} S_1 = \frac{\rho_{nh} \int_{F_x}^1 \int_{\mathbb{R}^1} \frac{1}{nh} K \left(\frac{x_i - X_i}{h} \right) \int_{F_{Z_1}}^1 \frac{\pm_i}{w(Z_1)} \int_{F_{Z_1}}^1 F(y|j|x) \int_{F_{Z_1}}^1$$

y, además considerar las funciones:

$$\begin{aligned} \mathcal{C}(u) &= E \int_{F_{Z_1}}^1 \frac{\pm_i}{w(Z_1)} \int_{F_{Z_1}}^1 F(y|j|x) \int_{F_{Z_1}}^1 \\ &= \int_{F_u}^1 [F(y|j|u) - F(y|j|x)] \end{aligned}$$

y

$$\begin{aligned} V(x) &= \text{Var} \int_{F_{Z_1}}^1 \frac{\pm_i}{w(Z_1)} \int_{F_{Z_1}}^1 F(y|j|x) \int_{F_{Z_1}}^1 \\ &= \int_{F_x}^1 (1 - 2F(y|j|x)) \int_0^Z \frac{dF(s|x)}{w(s)} + F^2(y|x) \int_0^Z \frac{dF(s|x)}{w(s)}; \end{aligned}$$

Teorema 3 (Eficiencia) Bajo las condiciones del teorema anterior, para cada $x \in I$ e $y \in \mathcal{Y}$; se tiene que $\mathcal{H}^2(y|x) = \mathcal{S}^2(y|x)$; siendo

$$\mathcal{S}^2(y|x) = \frac{(1 - F(y|j|x))^2 \int_{\mu_Z} \int_{\mathbb{R}^1} K^2(z) dz \int_0^{\mu_Z} \int_{\mathbb{R}^1} \frac{dH_1^n(s|x)}{C^2(s|x)}}{m^n(x)}$$

la varianza del estimador límite producto generalizado para $F(y|x)$, propuesto en Iglesias-Pérez y González-Manteiga (1999) para el caso general de datos censurados y truncados con covariables.

Demostración. Bajo las hipótesis de este trabajo se tiene que

$$dH_1^n(s|x) = \frac{w(s)}{1 - F_x} dF(s|x) \quad \text{y} \quad C(s|x) = \int_{F_x}^1 w(s) \int_{F_{Z_1}}^1 F(y|j|x) \int_{F_{Z_1}}^1$$

Ello conduce a la equivalencia entre $s^2(y, j, x) \geq \frac{3}{4} s^2(y, j, x) \geq 0$ y $D \geq 0$, siendo

$$D = \int_0^y \frac{(1 - F(y, j, x))^2}{w(s) [1 - F(s, j, x)]^2} ds - \int_0^y \frac{dF(s, j, x)}{w(s)} + F^2(y, j, x) \int_y^1 \frac{dF(s, j, x)}{w(s)} :$$

Finalmente, la prueba de que $D \geq 0$ es análoga a la demostración del Teorema 3.5 en de Uña-Álvarez (2003), sin más que tener en cuenta que $w(y)$ es una función no decreciente.

4. APLICACIÓN PRÁCTICA

Hemos aplicado el estimador condicional propuesto a los datos de supervivencia de diabetes mellitus en Dinamarca, recogidos en Green et al. (1981). Estos datos se discuten con detalle en Andersen et al. (1993) y un análisis condicional se muestra en Iglesias-Pérez (2003). Los análisis realizados son consistentes con investigaciones previas, pues muestran: una supervivencia relativa hombre-mujer inferior a uno (cuyo grado y evolución dependen de la edad de diagnóstico), así como una fuerte incidencia de la edad de diagnóstico en la supervivencia (el diagnóstico a edades tempranas se asocia con una supervivencia menor).

5. AGRADECIMIENTOS

Agradecemos al Profesor Per Kragh Andersen, del Departamento de Bioestadística de la Universidad de Copenhague, su cooperación al proporcionarnos los datos utilizados en este artículo, y al Dr. Anders Green la recopilación de dichos datos. Asimismo, el primer autor agradece el soporte económico de los proyectos PGIDIT02PXIA30003PR y BFM2002-03213.

6. REFERENCIAS

- Andersen, P. K., Borgan, Ø., Gill, R. D. y Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.
- Ashgharian, M., M'LAN, C. E. y Wolfson, D. B. (2002). "Length-biased sampling with right-censoring: an unconditional approach". *J. Amer. Statist. Assoc.* 97, 201-209.
- de Uña-Álvarez, J. (2003). "Empirical estimation under length-bias and type I censoring". *Reports in Statistics and OR*, nº 03-04, Universidad de Santiago de Compostela.
- Green, A., Hauge, M., Holm, N. V., y Rasch, L. L. (1981). "Epidemiological studies of diabetes mellitus in Denmark. II. A prevalence study based on insuline prescriptions". *Diabetologia* 20, 468-470.
- Iglesias-Pérez, M.C. y González-Manteiga, W. (1999). "Strong representation of a generalized product-limit estimator for truncated and censored data with some applications". *J. Nonparametric Statist.* 10, 213-244.
- Iglesias-Pérez, M.C. (2003). "Estimación de la función de distribución condicional en presencia de censura y truncamiento: una aplicación al estudio de la mortalidad en pacientes diabéticos". *Estadística Española*. Próximamente en el N° 153.
- Tsai, W. Y., Jewell, N. P. y Wang, M. C. (1987). "A note on the product-limit estimator under right censoring and left truncation". *Biometrika* 74, 883-886.
- Vardi, Y. (1982). "Nonparametric estimation in the presence of length bias". *Ann. Statist.* 10, 616-620.
- Wang, M.-C. (1991). "Nonparametric estimation from cross-sectional survival data". *J. Amer. Statist. Assoc.* 86, 130-143.
- Winter, B. B. y Földes, A. (1988). "A product-limit estimator for use with length-biased data". *Canad. J. Statist.* 16, 337-355.