QÜESTIIÓ, vol. 23, 2, p. 365-392, 1999

ESTUDIOS DE SUPERVIVENCIA CON DATOS NO OBSERVADOS. DIFICULTADES INHERENTES AL ENFOQUE PARAMÉTRICO

G. GÓMEZ*
C. SERRAT**
Universitat Politècnica de Catalunya

A partir de una muestra de datos de supervivencia que contiene valores no observados en las covariantes de interés, presentamos una metodología que permite extraer toda la información contenida en covariantes completamente observadas, que estén fuertemente correlacionadas con las citadas covariantes de interés. El enfoque utilizado es completamente paramétrico y se basa en el método de máxima verosimilitud. Mostramos las dificultades, tanto de índole práctica como filosófica, que aparecen en la especificación de la función de verosimilitud y en su optimización. Diseñamos una metodología que permite determinar en qué medida los estimadores resultantes dependen de la modelización del patrón de no respuesta y la implementamos sobre S-PLUS. Dicha metodología, a su vez, permite el estudio de la tipología de datos no observados y proporciona un análisis de sensibilidad de los resultados obtenidos. Ilustramos la metodología presentada y sus dificultades con una aplicación a una cohorte de pacientes con tuberculosis pulmonar infectados por el virus de la inmunodeficiencia humana.

Survival studies with non observed data. Difficulties concerning the parametric approach.

Palabras clave: Análisis de supervivencia, estudio de validación, máxima verosimilitud, modelos de datos incompletos, patrón de no respuesta aleatorio, patrón de no respuesta no ignorable.

Clasificación AMS (MSC 2000): 62F10, 62H99, 92C60.

Este trabajo ha sido parcialmente financiado por el proyecto DGICYT PB95–0776.

* Departament d'Estadística i Investigació Operativa. Universitat Politècnica de Catalunya. Pau Gargallo, 5. 08028 Barcelona. E-mail: ggg@eio.upc.es.

** Departament de Matemàtica Aplicada I. Universitat Politècnica de Catalunya. Av. Dr. Gregorio Marañón, 44–50. 08028 Barcelona. E-mail: carles@ma1.upc.es.

– Recibido en febrero de 1998.
– Aceptado en febrero de 1999.

1. INTRODUCCIÓN

El llamado *missing data problem*, y al que nos podemos referir como el problema de los datos no observados, es un problema clásico pero que continúa sin ser resuelto de forma satisfactoria. Dicha problemática contiene, en particular, aquellas situaciones en que la variable dependiente sí es observada totalmente, pero la información sobre las covariantes o variables independientes es incompleta. Dentro de ésta, y si uno se encuentra en un contexto de análisis de supervivencia, la variable dependiente, o tiempo de supervivencia, es parcialmente observada debido al llamado problema de censura.

La mayoría de las metodologías estadísticas existentes basan sus conclusiones en la suposición, no siempre explícita, de que los datos no observados son completamente aleatorios o, en el mejor de los casos, aleatorios (dichas definiciones se definirán de forma precisa en la sección 2 de este artículo). Un breve repaso a los diferentes enfoques, dentro del contexto del análisis de la supervivencia, así como a sus ventajas e inconvenientes, nos lleva a considerar al menos las siguientes cuatro posibilidades.

Un primer enfoque consiste en basar todos los análisis solamente en aquellos individuos que tienen todas las covariantes observadas. En este caso la inferencia conllevará resultados sesgados e inconsistentes, puesto que los individuos observados no tienen porqué ser representativos de toda la muestra parcialmente observada. Por otro lado, los estimadores basados en dichos análisis serán menos precisos debido a la reducción del tamaño de la muestra. Este tipo de análisis, desafortunadamente ampliamente usado, es obviamente práctico y sencillo de ser utilizado puesto que puede llevarse a cabo mediante el software existente. Además, aún en el caso en que la no observación de los datos se hubiera producido de forma completamente aleatoria, los estimadores deducidos por este método serían ineficientes.

Otra metodología posible consiste en la imputación de los valores no observados (Glynn, Laird y Rubin (1993), Efron (1994), Serrat y Gómez (1995)). El principal problema con dicho enfoque reside en la suposición de que los datos no observados siguen el mismo patrón que los observados. Obviamente, como dicha hipótesis no tiene porqué ser cierta, su uso conlleva un gran riesgo en la inferencia que se realiza e implica un sesgo evidente en los estimadores. Otro de los enfoques clásicos consiste en la modelización paramétrica del problema y en su resolución via la maximización de la función de verosimilitud. Este tipo de soluciones nace con los trabajos de Little y Rubin en la década de los 70; éstos proponen el método de la máxima verosimilitud con el objetivo de extraer la máxima información contenida en los datos observados (Little y Rubin (1987), Glynn, Laird y Rubin (1986)). La inferencia basada en este tipo de análisis goza de buenas propiedades asintóticas y es ampliamente usado; sin embargo, su implementación no es en absoluto sencilla y los estimadores resultantes dependen fuertemente del gran número de hipótesis que tienen que asumirse, lo que condiciona la credibilidad de los mismos. Últimamente, algunos autores, entre ellos Baker, proponen una metodo-

logía jerárquica basada también en el método de la máxima verosimilitud, que permite combinar distintos modelos de patrón de no respuesta y estudiar las variaciones de los estimadores bajo estos supuestos (Baker (1994)).

El enfoque semiparamétrico es una cuarta vía que permite modelar sólo aquello que es estrictamente necesario; en concreto, la relación entre la supervivencia y las covariantes, y la relación entre la supervivencia y el patrón de no respuesta. Los estimadores semiparamétricos son insesgados, consistentes y asintóticamente normales (Newey (1990), Rotnitzky y Wypij (1994), Robins, Rotnitzky y Zhao (1994)).

Este trabajo parte, por un lado, de las ideas desarrolladas por la Dra. Andrea Rotnitzky, de la Universidad de Harvard, en el curso de especialización en *Datos No Observados (Missing Data)* impartido en la Universitat Politècnica de Catalunya en julio de 1996 y, por otro, de las numerosas discusiones con ella mantenidas posteriormente. En el artículo, como punto previo a un análisis semiparamétrico, los autores abordan el problema de los datos no observados en un contexto de análisis de la supervivencia y desde una perspectiva totalmente paramétrica con un doble objetivo: a) mostrar las dificultades tanto de índole práctica como filosófica en la especificación de la función de verosimilitud y en la optimización de la misma y b) diseñar una metodología que permita determinar en qué medida los estimadores resultantes dependen del patrón de no respuesta.

La motivación de la metodología que pretendemos abordar surge de los estudios clínicos y epidemiológicos. En estos estudios disponemos de una cohorte de pacientes y queremos estudiar por un lado su supervivencia y por otro determinar aquellas variables que puedan ser predictoras de la misma. En particular, la colaboración, desde 1994, con los epidemiólogos del Institut Municipal de la Salut de Barcelona, ha motivado la necesidad de encontrar una metodología que permita «tratar» los valores no observados en las covariantes de interés.

El desarrollo del trabajo es como sigue. En la siguiente sección introducimos la notación así como las definiciones relevantes. En la sección 3 planteamos el problema y lo resolvemos paramétricamente. En la sección 4 presentamos como ilustración el análisis de los datos que dieron lugar a dichas reflexiones. El artículo finaliza con una discusión.

2. NOTACIÓN Y DEFINICIONES

Llamamos **vector de datos potenciales** $L = (L_1, L_2, \cdots, L_K)$ de un individuo arbitrario al vector de dimensión K formado por los datos observados y los datos no observados de dicho individuo. Asimismo, definimos el **vector de respuesta** de dicho individuo, y lo representamos por $R = (R_1, R_2, \cdots, R_K)$, como el vector cuya com-

ponente k-ésima ($k = 1, \cdots, K$) es igual a 1 si la variable k-ésima ha sido observada y es 0 en caso contrario.

Subdividimos el vector de datos potenciales L en dos subvectores $L_{(R)}$ y $L_{(\bar{R})}$ correspondientes a los datos observados y a los no observados, respectivamente. El subvector $L_{(R)}$ está formado por aquellas componentes l del vector L para las cuales $R_l = 1$, mientras que el subvector de las variables no observadas, $L_{(\bar{R})}$, consiste en aquellas componentes l del vector L para las cuales $R_l = 0$. Si por ejemplo $K = 5$ y $R = (1, 0, 1, 1, 0)$ entonces $L_{(R)} = (L_1, L_3, L_4)$ mientras que $L_{(\bar{R})} = (L_2, L_5)$.

Denotemos por $r = (r_1, r_2, \cdots, r_K)$, $r_k \in \{0, 1\}$, $k = 1, \cdots, K$, una realización del vector de respuesta de un individuo arbitrario. La probabilidad condicional de r dado el vector de datos potenciales L, la denotamos por $\pi_L(r) = P(R = r|L)$, y los diferentes tipos de procesos de no respuesta dependerán de los valores que ésta tome.

El proceso de no respuesta se denomina **completamente aleatorio** y se abrevia por MCAR (Missing Completely at Random) si, y sólo si, la probabilidad de una realización r es constante, es decir, si $\pi_L(r)$ no depende de L. El proceso es **aleatorio** o Missing at Random (MAR) si, y sólo si, la probabilidad de una realización r depende únicamente de las variables que han sido observadas, es decir, $\pi_L(r)$ depende sólo de $L_{(r)}$. Por último, el proceso de no respuesta es **no ignorable** si, y sólo si, la probabilidad condicional $\pi_L(r)$ depende del subvector de datos no observados $L_{(\bar{r})}$.

Si suponemos que las componentes del vector L siguen un orden establecido, el proceso de no respuesta se denomina **monótono** cuando la **no** observación de una variable implica la **no** observación de las siguientes; es decir, cuando $R_l = 0$ implica $R_m = 0$ para todo $m > l$.

En un estudio de supervivencia la variable de interés, que denotaremos por T, es usualmente el tiempo transcurrido desde un origen (*i.e.* aleatorización en un ensayo clínico, inicio de un tratamiento, etc.) hasta la realización de un cierto suceso (muerte, diagnóstico de SIDA, recaída de una enfermedad, etc.). Frecuentemente, en dichos estudios la realización de dicho suceso no es siempre observada, debido, bien a la finalización del estudio, bien a la pérdida de algunos de los individuos, dando lugar a la denominada censura por la derecha. Si denotamos por C el tiempo de censura, en realidad sólo observamos la variable $Y = min\{T, C\}$ y un indicador de censura $\delta = \mathbf{1}\{T \leq C\} = \mathbf{1}\{Y = T\}$ que indica si el dato ha sido o no censurado.

Para cada individuo, además de recoger los valores de (Y, δ), se recogen los valores de otras variables, llamadas covariantes, tales como indicadores sociodemográficos, variables de tipo clínico, resultados de análisis, etc. El objetivo es, en general, la modelización del tiempo de supervivencia en función de las covariantes de interés. Denotemos por X el vector formado por las covariantes de interés. Si X_* es una componente de X y V_* es otra covariante (no necesariamente componente de X), diremos que V_* es **sub-**

rogante de X_* si X_* y V_* están fuertemente correlacionadas. Por último, designemos por V el vector de covariantes subrogantes (no contenidas en X) para las componentes de interés.

Como es habitual en los análisis de supervivencia, supondremos que la distribución del tiempo de censura, C, es independiente de T, dado el vector de covariantes $(V^t, X^t)^t$.

Supongamos que disponemos de una muestra de individuos de tamaño n; de acuerdo con la notación introducida, para cada individuo i ($i = 1, \ldots, n$) representamos por $L_i = (Y_i, \delta_i, V_i^t, X_i^t)^t$ el vector de datos potenciales, por R_i el vector de respuesta y por $L_{(R_i)}$ y $L_{(\bar{R}_i)}$ los subvectores correspondientes a los datos observados y a los no observados, respectivamente. En nuestro estudio tenemos, por construcción del vector L_i, que $L_{i1} = Y_i$ y $L_{i2} = \delta_i$ y por consiguiente $R_{i1} = R_{i2} = 1$ para todo $i = 1, \ldots, n$.

3. PRUEBAS DE HIPÓTESIS PARA EL ESTUDIO DEL PROCESO DE NO RESPUESTA

En esta sección nos proponemos el estudio y validación del proceso de no respuesta subyacente en los datos. En primer lugar, planteamos pruebas sobre la hipótesis de que el proceso de no respuesta sea completamente aleatorio. En segundo lugar, y bajo una perspectiva totalmente paramétrica, introducimos un esquema jerárquico de validación a partir del cual realizamos un análisis de sensibilidad de la adecuación del modelo bajo los distintos patrones de no respuesta. Esta metodología permite, en particular, elucidar sobre la no ignorabilidad del proceso de no respuesta.

3.1. Validación del modelo MCAR

La validación de un patrón de no respuesta completamente aleatorio, dada la muestra observada, $(R_i, L_{(R_i)})_{i=1,2,\cdots,n}$, se basa en la comparación de las probabilidades $P(R_{ik} = 1)$, $i = 1, \cdots, n$ y $P(R_{ik} = 1|L_i), i = 1, \cdots, n$ para cada $k = 1, \cdots, K$. Para la resolución de la prueba de hipótesis

$$H_0 : \text{Proceso de no respuesta completamente aleatorio}$$
$$H_A : \text{Proceso de no respuesta aleatorio o no ignorable}$$

desarrollamos dos procedimientos en función de que el patrón de no respuesta sea monótono o que no lo sea.

Si el patrón de no respuesta es monótono la comparación anterior queda reducida a la comparación de las probabilidades $P(R_{ik} = 1|R_{i(k-1)} = 1)$, $i = 1, \cdots, n$ y $P(R_{ik} = 1|R_{i(k-1)} = 1, L_{i1}, \cdots, L_{i(k-1)})$, $i = 1, \cdots, n$ para cada $k = 1, \cdots, K$

(Apéndice, apartado a)). Si expresamos el *logit* de $P(R_{ik} = 1|R_{i(k-1)} = 1, L_{i1}, \cdots, L_{i(k-1)})$ como $\alpha_{k1} + \alpha_{k2}^t h_k(L_{i1}, \cdots, L_{i(k-1)})$ donde $h_k(L_{i1}, \cdots, L_{i(k-1)})$ es una función vectorial arbitraria de los datos $L_{i1}, \cdots, L_{i(k-1)}$, la prueba de hipótesis H_0 contra H_A puede plantearse como las siguientes K pruebas simultáneas:

$$\begin{aligned} H_{0k} &: logit(P(R_{ik} = 1|R_{i(k-1)} = 1, L_{i1}, \cdots, L_{i(k-1)})) = \alpha_{k1} \\ H_{Ak} &: logit(P(R_{ik} = 1|R_{i(k-1)} = 1, L_{i1}, \cdots, L_{i(k-1)})) = \\ &\quad = \alpha_{k1} + \alpha_{k2}^t h_k(L_{i1}, \cdots, L_{i(k-1)}) \qquad\qquad k = 1, \cdots, K. \end{aligned}$$

Para cada k, $k = 1, \cdots, K$, el estadístico basado en la razón de verosimilitud de H_{0k} contra H_{Ak} nos proporciona un p-valor, p_k. En el apartado b) del Apéndice demostramos que si los datos son MCAR, $i.e.$ las hipótesis $H_{0k}, k = 1, \cdots, K$, son todas verdaderas, entonces los p-valores resultantes siguen una distribución uniforme en (0,1) y son independientes entre sí.

Para la interpretación conjunta de los K p-valores resultantes, $p_1, \cdots, p_K$, utilizamos el estadístico combinado $S = -2\sum_{k=1}^{K} \log p_k$, que sigue bajo H_0 una distribución χ^2 con $2K$ grados de libertad.

Si el patrón de no respuesta es no monótono la comparación de H_0 versus H_A debe resolverse a partir de la comparación, para cada $i = 1, \cdots, n$ y para cada valor de k, $k = 1, \cdots, K$, de las probabilidades $P(R_{ik} = 1)$ y $P(R_{ik} = 1|$ datos observados para $k' \neq k)$. En este caso los p-valores obtenidos no son independientes y en el diseño de las correspondientes pruebas de hipótesis se deben utilizar técnicas de inferencia simultánea (Miller, 1980), como por ejemplo el estadístico t de Bonferroni. En general, estas técnicas son más conservadoras y en consecuencia reducirán la potencia de las pruebas de hipótesis resultantes. En la sección 4.2 presentamos detalladamente la aplicación de esta metodología.

3.2. Planteamiento paramétrico del problema

Planteamos el problema de la estimación del modelo de supervivencia de T desde una perspectiva completamente paramétrica y a partir de la muestra de datos potenciales $L_i = (Y_i, \delta_i, V_i^t, X_i^t)^t$, $i = 1, 2, \cdots, n$, donde X es un vector de covariantes parcialmente observadas y V es un vector de covariantes completamente observadas. Dicha perspectiva nos va a permitir introducir la modelización de las probabilidades de no respuesta y, en consecuencia, estudiar la bondad del ajuste para dichas modelizaciones. Dicha estimación se lleva a cabo mediante un análisis de máxima verosimilitud. Con el objetivo de especificar la función de verosimilitud, L, para dicha muestra, denotamos por $f_c(l; \theta)$ la función de densidad de los datos completos, y por $P(R_i = r|L_i; \psi)$ las

probabilidades de observar exactamente ciertas componentes de L_i. Nótese que dichas funciones dependen, respectivamente, de sendos parámetros θ y ψ, que se suponen de variación independiente.

La contribución del individuo i-ésimo a la función de verosimilitud $L(\theta, \psi)$ es: $f_c(L_i; \theta) \cdot P(R_i = \mathbf{1}|L_i; \psi)$ si el individuo ha sido completamente observado (*i.e.*, $R_i = \mathbf{1}$), y $\int f_c(L_i; \theta) \cdot P(R_i = r|L_i; \psi) dL_{(\bar{r})i}$ si el individuo ha sido parcialmente observado y su vector de respuesta es $R_i = r$ con $r \neq \mathbf{1}$. Esta segunda expresión corresponde a la marginalización de la primera respecto a los datos no observados. Así, la función de verosimilitud $L(\theta, \psi)$ a partir de los datos observados es

$$\begin{aligned} L(\theta, \psi) &= \prod_{i=1}^{n} \left\{ [f_c(L_i; \theta) \cdot P(R_i = \mathbf{1}|L_i; \psi)]^{I(R_i = \mathbf{1})} \right. \\ &\quad \left. \prod_{r \neq \mathbf{1}} \left[\int f_c(L_i; \theta) \cdot P(R_i = r|L_i; \psi) dL_{(\bar{r})i} \right]^{I(R_i = r)} \right\}. \end{aligned}$$

En el apartado c) del Apéndice demostramos que, cuando el patrón de no respuesta es MCAR o MAR, es decir, cuando las probabilidades $P(R_i = r|L_i)$ no dependen de $L_{(\bar{r})i}$, dichas probabilidades se pueden factorizar en la expresión de la verosimilitud, $L(\theta, \psi)$, y, por consiguiente, la estimación máximo verosímil del parámetro θ es independiente del patrón de no respuesta utilizado.

En nuestro estudio la función de densidad de los datos, $f_c(L_i; \theta)$, toma la expresión

$$\begin{aligned} f_c(L_i; \theta) &= f_c((Y_i, \delta_i, V_i^t, X_i^t)^t; \theta) = \\ &= f_c(X_i; \theta) \cdot f_c(Y_i, \delta_i|X_i; \theta) \cdot f_c(V_i|Y_i, \delta_i, X_i; \theta). \end{aligned}$$

La modelización paramétrica del problema exige la correcta especificación de: a) la distribución de las covariantes de interés, X; b) la distribución de los tiempos observados, Y, condicionada a las covariantes X; c) la distribución de las covariantes subrogantes, V, condicionadas a Y y a X, y d) las probabilidades de respuesta condicionadas a los datos potenciales. Hemos de tener en cuenta que dichas especificaciones pueden ser hasta cierta medida arbitrarias y que, además, ninguna de las citadas distribuciones es validable a partir de los datos observados.

Supongamos en lo que sigue que el vector de covariantes del i–ésimo individuo, X_i, está formado por p variables aleatorias discretas, $X_{i1}, X_{i2}, \cdots, X_{ip}$ y que el vector V_i de covariantes subrogantes contiene asimismo q variables aleatorias discretas, $V_{i1}, V_{i2}, \cdots, V_{iq}$.

La especificación de las funciones de densidad $f_c(X_i; \theta)$ y $f_c(V_i|Y_i, \delta_i, X_i; \theta)$, así como la determinación de las probabilidades $P(R_i = r|L_i; \psi)$, puede hacerse en términos de los logaritmos de las *odds ratio* condicionadas de cada categoría respecto al grupo de

referencia (regresiones logísticas condicionadas, en el caso binario), es decir, dando una modelización para cada una de las siguientes expresiones:

$$\log \frac{P(X_{ij}=k|X_{i1},\cdots,X_{i(j-1)})}{P(X_{ij}=0|X_{i1},\cdots,X_{i(j-1)})} \quad j=1,\cdots,p \quad k\neq 0,$$

$$\log \frac{P(V_{ij}=k|V_{i1},\cdots,V_{i(j-1)})}{P(V_{ij}=0|V_{i1},\cdots,V_{i(j-1)})} \quad j=1,\cdots,q \quad k\neq 0 \quad \text{y}$$

$$logit(P(R_{ij}=1|R_{i1},\cdots,R_{i(j-1)})) \quad j=1,\cdots,p.$$

Como veremos en la ilustración de la siguiente sección, el uso de distintos modelos en la expresión de las probabilidades $P(R_{ij}=1|R_{i1},\cdots,R_{i(j-1)})$, $j=1,\cdots,p$, en función de los datos L_i permite analizar la sensibilidad de los estimadores al patrón de no respuesta utilizado.

La modelización de los tiempos de supervivencia y su relación con las covariantes X, queda restringida a la especificación de la función de densidad $f_c(Y_i,\delta_i|X_i;\theta)$. En este caso, pueden ser útiles análisis previos basados en la submuestra completamente observada. Sin embargo, una vez más, estas suposiciones no se pueden validar a partir de los datos observados.

Una limitación añadida a las anteriores modelizaciones es el crecimiento rápido de la dimensión de los parámetros θ y ψ. Este hecho puede repercutir, de manera importante, en la ejecución de las correspondientes implementaciones, a la vez que reduce el tamaño muestral relativo. Por ejemplo, es fácil calcular que en el simple supuesto de que todas las covariantes fueran binarias y usando sólo modelos lineales que no incluyesen interacciones, tendríamos

$$\begin{aligned} dim(\theta) &= (2^p-1)+(p+1)+(2^q-1)(p+3), \\ dim(\psi) &= (2^p-1)(p+q+3), \end{aligned}$$

y para una situación sencilla en la que $p=q=3$, $dim(\theta)=53$, $dim(\psi)=63$ y por consiguiente tendríamos un total de 116 parámetros en la función de verosimilitud.

4. ILUSTRACIÓN

Como ilustración de la metodología de la sección anterior presentamos una aplicación en un estudio de supervivencia en una cohorte de pacientes con tuberculosis pulmonar infectados por el virus de la inmunodeficiencia humana (VIH). Dicha aplicación forma parte de la colaboración que los autores mantienen con el Servicio de Epidemiología del Institut Municipal de la Salut de Barcelona. Los datos proceden de distintos registros de pacientes del Programa de Prevención y Control de la Tuberculosis. Entre los objetivos

epidemiológicos del citado Programa destacan: a) el estudio de la progresión del SIDA en pacientes tuberculosos y b) la determinación de indicadores de supervivencia en pacientes VIH+ y con tuberculosis.

4.1. Material y métodos

Se dispone de una muestra de 418 pacientes VIH+ con tuberculosis pulmonar residentes en Barcelona que fueron diagnosticados con tuberculosis en el período 1992–1994. El cierre del estudio se realizó en fecha 30 de septiembre de 1995. El tiempo de supervivencia de interés es el transcurrido entre el diagnóstico de la enfermedad (con el consiguiente immediato inicio del tratamiento antituberculosis) y la muerte. Para cada individuo de la muestra se dispone, potencialmente, de las siguientes variables de tipo sociológico y clínico, recogidas a su entrada en el estudio: sexo, edad, distrito de residencia, antecedentes de prisión, seguimiento anterior de un tratamiento antituberculosis, pertenencia a un grupo de prácticas de riesgo, localización de la tuberculosis, resultados radiológicos, resultados bacteriológicos, porcentajes de subpoblaciones linfocitarias T-CD4+ y T-CD8+, resultado de la prueba de la tuberculina, etc.

Estudios previos (Caylà *et al.*, 1993; Serrat y Gómez, 1995) realizados con datos completamente observados demuestran que para la estimación del mencionado tiempo de supervivencia las covariantes de interés son el porcentaje de T-CD4+ (y en particular su dicotomización en un nivel bajo y uno alto de inmunodepresión) y el resultado de la prueba de la tuberculina. Nos referiremos a estas variables por CD4 y PPD, respectivamente, y serán las componentes del vector X introducido en la sección 2. Distinguiremos los pacientes por el nivel de inmunodepresión según si la variable CD4 toma valores superiores al 14 % o no. La variable PPD toma valor 1 si el resultado de la prueba de la tuberculina es positivo, y 0 en caso contrario. El problema metodológico viene motivado por el hecho de que en nuestra muestra dichas variables presentan un 37.5 % y 50.5 % de valores no observados, respectivamente. Más concretamente, sólo se dispone de ambas en un 31.3 % de los individuos de la muestra y hay un 19.1 % de la muestra que no tiene recogido el valor de ninguna de las dos variables.

El objetivo es determinar el carácter predictivo de los indicadores CD4 y PPD haciendo uso de toda la información contenida en la muestra y, en particular, estudiar el resultado de la prueba de la tuberculina como medida de calidad complementaria al nivel de respuesta inmunológica que proporciona el recuento de T-CD4+. Para todo ello utilizaremos la metodología descrita en la sección anterior.

Después de estudiar, conjuntamente con el equipo de epidemiólogos, qué variables podrían proporcionar información cualitativa sobre el patrón de no respuesta o sobre las variables CD4 y PPD, se han elegido: a) el haber seguido un tratamiento anterior (TR: 1=Sí y 2=No); b) la radiología (RA: 0 = Normal, 1 = Anormal con patrón cavitario y 2

= Anormal sin patrón cavitario), y c) la bacteriología (BA: 0 = Negativa, 1 = Positiva y 2 = Positiva con cultivo bacteriológico) como covariantes subrogantes, V, de las covariantes de interés, X. Así pues, de acuerdo con la notación introducida, nuestro vector de datos es:

$$L_i = (Y_i, \delta_i, TR_i, RA_i, BA_i, \mathbf{1}\{CD4_i > 14\}, PPD_i)^t.$$

Para simplificar la notación omitiremos, siempre que no se preste a confusión, el subíndice i.

Observemos que, dado que las covariantes subrogantes son completamente observadas, $R_3 = R_4 = R_5 \equiv 1$. Así pues, el vector de observaciones $r \in \{0,1\}^7$, tomará únicamente los valores (1,1,1,1,1,1,1), (1,1,1,1,1,1,0), (1,1,1,1,1,0,1) y (1,1,1,1,1,0,0).

4.2. Validación del modelo MCAR

Para la validación del modelo MCAR aplicamos la metodología presentada en la sección 3.1. Como los datos no responden a un patrón de no respuesta monótono, hemos de comparar la probabilidad $P(R_{CD4} = 1)$ con $P(R_{CD4} = 1|Y, \delta, TR, RA, BA, PPD)$, y $P(R_{PPD} = 1)$ con $P(R_{PPD} = 1|Y, \delta, TR, RA, BA, \mathbf{1}\{CD4 > 14\})$.

Para el primer caso, si modelamos la probabilidad de respuesta a la variable $CD4$ como función de las otras variables observadas según el modelo logístico

$$\begin{aligned} &logit(P(R_{CD4} = 1|Y, \delta, TR, RA, BA, PPD)) = \\ &= \alpha_0 + \alpha_1 Y + \alpha_2 \delta + \alpha_3 TR + \alpha_4 RA + \alpha_5 BA + \alpha_6 PPD, \end{aligned}$$

la comparación entre las probabilidades $P(R_{CD4} = 1)$ y $P(R_{CD4} = 1|Y, \delta, TR, RA, BA, PPD)$ es equivalente a la siguiente prueba de hipótesis

$$\begin{aligned} H_0 &: \alpha_i = 0 \quad \forall i \in \{1, \cdots, 6\} \\ H_A &: \exists i \in \{1, \cdots, 6\} | \alpha_i \neq 0. \end{aligned}$$

Si es verdadera la hipótesis nula, la probabilidad de respuesta a la variable $CD4$ no dependerá de las otras variables observadas y su contribución en la prueba de hipótesis MCAR del proceso de no respuesta será en el sentido de no rechazar este supuesto. En caso contrario, de no ser la hipótesis nula cierta, tendremos evidencia para rechazar la hipótesis de no respuesta MCAR.

Análogamente, aplicamos también esta metodología para la probabilidad de respuesta a la variable PPD, con el modelo

$$\begin{aligned} &logit(P(R_{PPD} = 1|Y, \delta, TR, RA, BA, \mathbf{1}\{CD4 > 14\})) = \\ &= \beta_0 + \beta_1 Y + \beta_2 \delta + \beta_3 TR + \beta_4 RA + \beta_5 BA + \beta_6 \mathbf{1}\{CD4 > 14\}. \end{aligned}$$

Para combinar ambas pruebas utilizamos la corrección de Bonferroni. Esta corrección consiste en utilizar como nivel de significación en cada prueba de hipótesis parcial el nivel de significación global dividido por el número de pruebas que se desean combinar. El criterio de decisión que se utiliza es el siguiente: se rechaza la hipótesis nula cuando alguna prueba parcial rechaza su correspondiente hipótesis nula. Observemos que, por una parte, puede parecer más immediato rechazar la hipótesis nula por el simple hecho de utilizar más de una prueba; ahora bien, el hecho de utilizar una fracción del nivel de significación en cada prueba parcial exige más evidencia en contra de la respectiva hipótesis nula parcial. El efecto combinado es una prueba, en general, más conservadora.

En nuestro caso, las pruebas de hipótesis para las variables R_{CD4} y R_{PPD} tienen un p-valor 0.03766 y 0.1733, respectivamente. Si utilizamos un nivel de significación global del $5\,\%$, ambos resultados no resultan significativos (son mayores de 0.025) con lo que en ambas variables no se detecta evidencia en contra de la hipótesis nula. En consecuencia, no podemos rechazar la hipótesis que el proceso de no respuesta sea MCAR.

4.3. Planteamiento paramétrico del problema

Siguiendo los pasos indicados en la sección 3.2, la contribución de un individuo a la función de verosimilitud $L(\theta, \psi)$ se puede calcular a partir de las cuatro expresiones que a continuación se detallan.

a) Para las distribuciones de las covariantes de interés, $CD4$ y PPD, se han usado modelos logísticos para la probabilidad de $\mathbf{1}\{CD4 > 14\}$ y para las probabilidades de $PPD = 1$ condicionadas a los valores de $\mathbf{1}\{CD4 > 14\}$, es decir

$$\begin{aligned} logit(P(\mathbf{1}\{CD4 > 14\} = 1)) &= \alpha_1, \\ logit(P(PPD = 1|\mathbf{1}\{CD4 > 14\} = 1)) &= \alpha_2 \quad \text{y} \\ logit(P(PPD = 1|\mathbf{1}\{CD4 > 14\} = 0)) &= \alpha_3. \end{aligned}$$

b) En un estudio inicial realizado sobre la misma cohorte de pacientes (Serrat *et al.*, 1998) se observó que el tiempo de supervivencia podía ser modelado satisfactoriamente mediante una distribución de Weibull con covariantes CD4 y PPD. Esta misma modelización es la utilizada en esta aplicación y por tanto la función $f_c(Y, \delta|X; \theta)$ se expresa como

$$f_c(Y, \delta|CD4, PPD; \theta) = \left(\frac{1}{\sigma Y} (e^{-\beta} Y) \frac{1}{\sigma} \right)^{\delta} \cdot e^{-(e^{-\beta} Y)^{\frac{1}{\sigma}}},$$

donde $\beta = \beta_0 + \beta_1 \mathbf{1}\{CD4 > 14\} + \beta_2 PPD$ y $\sigma = \sigma_0 + \sigma_1 \mathbf{1}\{CD4 > 14\} + \sigma_2 PPD$.

c) Las distribuciones de TR, RA y BA dados $Y, \delta, CD4, PPD$ se modelizan a partir de los logaritmos de las *odds ratio* respecto del grupo de referencia, condicionadas a las variables previas. Si por $\lambda_{\cdot}$ denotamos de manera genérica una de las siguientes *odds ratio*

$$\begin{aligned}
\lambda_1 &= \frac{P(TR=1)}{P(TR=0)}, \\
\lambda_{2ij} &= \frac{P(RA=j|TR=i)}{P(RA=0|TR=i)} \quad i=0,1 \quad j=1,2, \\
\lambda_{3ijk} &= \frac{P(BA=k|TR=i,RA=j)}{P(BA=0|TR=i,RA=j)} \quad i=0,1 \quad j=0,1,2 \quad k=1,2,
\end{aligned}$$

entonces su logaritmo, $\log(\lambda_{\cdot})$, queda especificado por

$$\log(\lambda_{\cdot}) = \gamma_{,0} + \gamma_{,1}\log(Y) + \gamma_{,2}\delta + \gamma_{,3}\log(Y)\delta + \gamma_{,4}\mathbf{1}\{CD4 > 14\} + \gamma_{,5}PPD,$$

y está en función de las variables Y, δ y X, de algunas posibles interacciones entre ellas, y del parámetro $(\gamma_{,0}, \cdots, \gamma_{,5})$. Obsérvese que para reducir el efecto debido a los valores extremos en los tiempos de supervivencia observados, hemos reescalado los tiempos a escala logarítmica.

d) Los distintos patrones de no respuesta los diseñamos a partir de la modelización de las probabilidades $P(R_{ij} = 1|R_{i1}, \cdots, R_{i(j-1)})$ en términos de las covariantes V y X. Definimos modelos logísticos para la probabilidad de observación de la variable $CD4$ y para las probabilidades de observación de la variable PPD condicionadas a los valores de la variable $CD4$ de acuerdo con el siguiente esquema:

$$\begin{aligned}
logit(P(R_{CD4}=1)) &= \alpha_{10}+ \\
&+ \alpha_{11}TR + \alpha_{12}RA1 + \alpha_{13}RA2 + \alpha_{14}BA1 + \alpha_{15}BA2 \\
&+ \alpha_{16}\mathbf{1}\{CD4>14\} + \alpha_{17}PPD, \\
logit(P(R_{PPD}=1|R_{CD4}=1)) &= \alpha_{20}+ \\
&+ \alpha_{21}TR + \alpha_{22}RA1 + \alpha_{23}RA2 + \alpha_{24}BA1 + \alpha_{25}BA2 \\
&+ \alpha_{26}\mathbf{1}\{CD4>14\} + \alpha_{27}PPD \quad \text{y} \\
logit(P(R_{PPD}=1|R_{CD4}=0)) &= \alpha_{30}+ \\
&+ \alpha_{31}TR + \alpha_{32}RA1 + \alpha_{33}RA2 + \alpha_{34}BA1 + \alpha_{35}BA2 \\
&+ \alpha_{36}\mathbf{1}\{CD4>14\} + \alpha_{37}PPD,
\end{aligned}$$

donde $RAi = \mathbf{1}\{RA = i\}, i = 1, 2$ y $BAi = \mathbf{1}\{BA = i\}, i = 1, 2$ denotan las variables binarias que recogen los efectos de las categorías en radiología y bacteriología, respectivamente.

El parámetro θ de la densidad $f_c(L_i;\theta)$ en la verosimilitud $L(\theta,\psi)$ es

$$\theta = (\alpha_1, \alpha_2, \alpha_3, \beta_0, \beta_1, \beta_2, \sigma_0, \sigma_1, \sigma_2, \gamma_{,0}, \cdots, \gamma_{,5})$$

y tiene dimensión 111. El parámetro complementario resultante de la modelización de las probabilidades de respuesta es

$$\psi = (\alpha_{10}, \cdots, \alpha_{37})$$

y tiene dimensión 24. Observemos que el parámetro de interés, a efectos de estimación, está formado únicamente por las componentes $(\beta_0, \beta_1, \beta_2, \sigma_0, \sigma_1, \sigma_2)$ del vector θ y es de dimensión 6.

La jerarquización de las probabilidades del apartado d) permite simular, de forma anidada, los distintos patrones de no respuesta definidos en la sección 2. En nuestro estudio optimizamos la función de verosimilitud para los siguientes cinco casos:

$$\begin{aligned}
\alpha_{ij} = 0 \quad i = 1,2,3 \quad j = 1,\cdots,7 \quad &\Rightarrow \quad \text{MCAR} \\
\alpha_{ij} = 0 \quad i = 1,2,3 \quad j = 6,7 \quad &\Rightarrow \quad \text{MAR} \\
\alpha_{ij} = 0 \quad i = 1,2,3 \quad j = 7 \quad &\Rightarrow \quad \text{No ignorable (1r caso) NI1} \\
\alpha_{ij} = 0 \quad i = 1,2,3 \quad j = 6 \quad &\Rightarrow \quad \text{No ignorable (2o caso) NI2} \\
\text{Sin restricciones sobre los valores de } \alpha_{ij} \quad &\Rightarrow \quad \text{No ignorable (3r caso) NI3}
\end{aligned}$$

La tabla 1 muestra la estimación de los cuartiles relativos del grupo tuberculín positivo ($PPD = 1$) respecto del grupo tuberculín negativo ($PPD = 0$), bajo los distintos supuestos de covariantes subrogantes y, en cada uno de ellos, considerando los distintos patrones de no respuesta citados.

El uso de otros modelos basados en las mismas covariantes permitiría llevar a cabo un análisis de sensibilidad más exhaustivo de los estimadores resultantes y podría dar una respuesta global al problema en función del patrón de no respuesta subyacente en los datos.

La implementación de esta metodología se ha realizado en S-PLUS y ejecutado en un ordenador PC-Pentium Pro, 200 Mhz con 32 Mb RAM, en entorno Windows 95.

4.4. Interpretación de los resultados

En cuanto a la validación del modelo MCAR, hemos visto que, si bien existen técnicas eficientes para su validación bajo un patrón de no respuesta monótono, las técnicas existentes para el caso no monótono no gozan de potencia suficiente. En otros términos, si se utiliza dicha metodología se hace necesaria una mayor evidencia en los datos para

rechazar la hipótesis MCAR sobre el modelo de no respuesta. En nuestro caso sería necesario utilizar un nivel de confianza máximo del 92.4 $\%$ para rechazar la hipótesis MCAR.

Tabla 1. Estimación de cuartiles relativos para el grupo tuberculín positivo respecto del grupo tuberculín negativo, bajo los distintos supuestos de covariantes subrogantes (V) y modelos de patrón de no respuesta (M_i).

Covariantes Subrogantes V	Patrón de no respuesta M_i	Primer cuartil	Mediana	Tercer cuartil	*Deviance* D_{M_i}	Número de parámetros n_i
Ninguna	MCAR=MAR	2.983	1.630	1.012	4102.055	12
	NI1	2.293	1.384	0.929	4092.145	15
	NI2	2.675	1.479	0.927	4092.993	15
	NI3	1.784	1.125	0.782	4085.622	18
TR	MCAR/MAR	2.674	1.489	0.939	4522.524/4513.648	18/21
	NI1	1.973	1.270	0.897	4504.61	24
	NI2	2.640	1.474	0.931	4504.236	24
	NI3	1.780	1.121	0.779	4496.693	27
RA	MCAR/MAR	2.963	1.592	0.976	4688.37/4683.306	24/30
	NI1	2.251	1.341	0.892	4672.024	33
	NI2	2.652	1.483	0.938	4674.658	33
	NI3	1.178	1.122	0.781	4664.661	36
BA	MCAR/MAR	2.912	1.644	1.048	4962.181/4957.538	24/30
	NI1	2.262	1.391	0.948	4947.412	33
	NI2	2.657	1.484	0.937	4947.574	33
	NI3	1.776	1.124	0.783	4939.764	36
TR, RA	MCAR/MAR	2.751	1.496	0.926	5102.458/5087.667	42/51
	NI1	2.013	1.265	0.877	5077.458	54
	NI2	2.623	1.479	0.942	5077.824	54
	NI3	2.039	1.284	0.892	5070.985	57
TR, BA	MCAR/MAR	2.102	1.264	0.847	5356.891/5343.023	42/51
	NI1	2.164	1.229	0.787	5334.777	54
	NI2	2.182	1.286	0.848	5340.131	54
	NI3	1.245	1.423	1.580	5326.711	57
RA, BA	MCAR/MAR	2.890	1.641	1.050	5513.294/5504.302	60/72
	NI1	2.316	1.426	0.974	5492.744	75
	NI2	2.740	1.524	0.960	5494.363	75
	NI3	2.273	1.415	0.973	5487.814	78
TR, RA, BA	MCAR/MAR	2.739	1.451	0.880	5888.679/5863.155	114/129
	NI1	2.691	1.442	0.881	5853.455	132
	NI2	2.773	1.456	0.877	5854.811	132
	NI3	1.898	1.177	0.807	5838.814	135

Tabla 2. Comparación de los ajustes obtenidos en la estimación paramétrica, bajo distintos supuestos de covariantes subrogantes (V) y modelos de patrón de no respuesta (M_i). $\Delta_{M_i M_j} = (-2\log L_{M_i}) - (-2\log L_{M_j}) = D_{M_i} - D_{M_j}$.
(*) Resultado significativo al 95 %,
(**) Resultado significativo al 99 %

Covariantes Subrogantes V	Patrón de no respuesta M_i	Patrón de no respuesta M_j	$\Delta_{M_i\, M_j}$	Grados de libertad $n_j - n_i$	$p-$value
Ninguna	MCAR=MAR	NI1	9.91	3	0.019(*)
	MCAR=MAR	NI2	9.062	3	0.029(*)
	NI1	NI3	6.523	3	0.089(*)
	NI2	NI3	7.371	3	0.061(*)
TR	MCAR	MAR	8.876	3	0.031(*)
	MAR	NI1	9.038	3	0.029(*)
	MAR	NI2	9.412	3	0.024(*)
	NI1	NI3	7.917	3	0.048(*)
	NI2	NI3	7.543	3	0.057
RA	MCAR	MAR	5.064	6	0.536
	MAR	NI1	11.282	3	0.010(*)
	MAR	NI2	8.648	3	0.034(*)
	NI1	NI3	7.363	3	0.061
	NI2	NI3	9.997	3	0.019(*)
BA	MCAR	MAR	4.643	6	0.59
	MAR	NI1	10.126	3	0.018(*)
	MAR	NI2	9.964	3	0.019(*)
	NI1	NI3	7.648	3	0.054
	NI2	NI3	7.81	3	0.050
TR, RA	MCAR	MAR	14.791	9	0.097
	MAR	NI1	10.209	3	0.017(*)
	MAR	NI2	9.843	3	0.02(*)
	NI1	NI3	6.473	3	0.091
	NI2	NI3	6.839	3	0.077
TR, BA	MCAR	MAR	13.868	9	0.127
	MAR	NI1	8.246	3	0.041(*)
	MAR	NI2	2.892	3	0.409
	NI1	NI3	8.066	3	0.045(*)
	NI2	NI3	13.42	3	0.004(**)
RA, BA	MCAR	MAR	8.992	12	0.704
	MAR	NI1	11.558	3	0.009(**)
	MAR	NI2	9.939	3	0.019(*)
	NI1	NI3	4.93	3	0.177
	NI2	NI3	6.549	3	0.088
TR, RA, BA	MCAR	MAR	25.524	15	0.043(*)
	MAR	NI1	9.7	3	0.021(*)
	MAR	NI2	8.344	3	0.039(*)
	NI1	NI3	14.641	3	0.002(**)
	NI2	NI3	15.997	3	0.001(**)

La sensibilidad de la estimación de los parámetros a la modelización del patrón de no respuesta, como queda reflejado en la tabla 1, demuestra que el patrón de no respuesta no es MCAR e ilustra la baja potencia de dicha metodología.

Respecto a la estimación paramétrica, para la comparación de los resultados obtenidos, para un cierto subconjunto de covariantes subrogantes, según los distintos modelos anidados del patrón de no respuesta (MCAR a NI3), utilizaremos el estadístico resultante de la diferencia entre las *deviances* ($-2\log L$) de las dos modelizaciones.

Si denotamos por D_{M_i} la *deviance* obtenida en la modelización M_i, $i = 1, \cdots, 5$, el estadístico $\Delta_{M_i M_j} = D_{M_i} - D_{M_j}$ sigue una distribución $\chi^2_{n_j - n_i}$ donde n_i y n_j $(n_i < n_j)$ son el número de parámetros a estimar bajo cada una de las dos modelizaciones anidadas M_i y M_j, respectivamente. La tabla 2 presenta los p-valores obtenidos en la comparación de los modelos paramétricos resultantes, bajo distintos supuestos de covariantes subrogantes (V) y modelos de patrón de no respuesta (M_i).

Del análisis de las tablas 1 y 2 podemos concluir:

1. En nuestra aplicación, en general, el modelo MAR no resulta significativo respecto al correspondiente MCAR. La interpretación inmediata es pues que la probabilidad de observación de las variables $CD4$ y PPD poco depende de los valores observados en las variables subrogantes.

Concretamente, sólo obtenemos diferencias significativas cuando utilizamos como variable subrogante el haber seguido recientemente tratamiento antituberculosis. En este caso las estimaciones de los coeficientes correspondientes a la variable TR resultan de signo positivo; esto nos permite concluir, como parece lógico, que aquellos pacientes que tienen antecedentes de tratamiento antituberculosis tienen mayor probabilidad de tener informadas las variables $CD4$ y PPD.

2. En todos los supuestos, los modelos No Ignorables resultan significativos respecto a los MCAR y MAR. Esto nos indica que efectivamente la probabilidad de observación de las variables CD4 y PPD depende de los valores potenciales de dichas variables. Observemos que este resultado es imposible de obtener a partir de pruebas de hipótesis basadas en los datos observados (Sección 3.1)

3. La positividad en la prueba de la tuberculina es un mejor pronóstico de supervivencia *a corto plazo*; sin embargo, a *largo plazo* la supervivencia es peor. Para ilustrar este hecho presentamos en la figura 1 la estimación de la función de supervivencia para el grupo tuberculín positivo y para el grupo tuberculín negativo, según el nivel de inmunodepresión, cuando utilizamos como covariantes subrogantes las variables TR y RA, y como patrón de no respuesta el NI2. Las estimaciones correspondientes a los parámetros β y σ de la distribución de Weibull son $\beta = 6{,}878 + 1{,}362 \cdot \mathbf{1}\{CD4 > 14\} + 0{,}153 \cdot PPD$ y $\sigma = 1{,}426 + 0{,}244 \cdot \mathbf{1}\{CD4 > 14\} - 0{,}651$

$\cdot PPD$, respectivamente. Esto nos permite estimar la distribución de cuartiles para el grupo más inmunodeprimido en 164, 576 y 1547 días si el resultado de la prueba de la tuberculina es negativo o en 431, 851 y 1457 días si el resultado de dicha prueba es positivo. Análogamente, para el grupo menos inmunodeprimido, podemos estimar los cuartiles en 473, 2055 y 6539 días o 1241, 3040 y 6160 días, según el resultado de la prueba de la tuberculina. Efectivamente a corto y medio plazo (hasta aproximadamente el percentil del 72 %) la supervivencia del grupo tuberculín positivo es mejor −independientemente del nivel de inmunodepresión− que la del grupo tuberculín negativo.

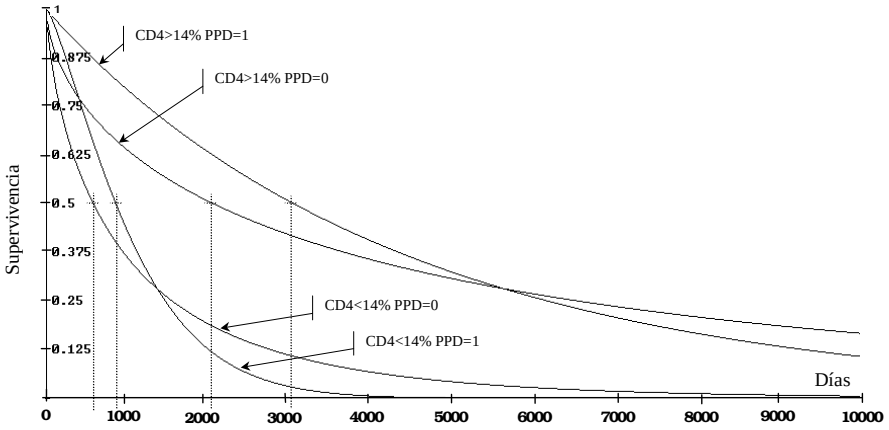

Figura 1. Estimación de la función de supervivencia para el grupo tuberculín positivo ($PPD = 1$) y para el grupo tuberculín negativo ($PPD = 0$), según el nivel de inmunodepresión (alto $\equiv$ $CD4 \leqslant 14\,\%$, bajo $\equiv CD4 > 14\,\%$), cuando se utilizan las covariantes TR y RA como subrogantes y el modelo NI2 como patrón de no respuesta.

5. DISCUSIÓN

La metodología presentada adolece de ciertas limitaciones que hace falta conocer y que condicionan su aplicabilidad. Por un lado, es realmente difícil especificar correctamente tanto los modelos usados para las distintas funciones de densidad, como para las probabilidades de respuesta. Una consecuencia inmediata de dicha dificultad es la introducción de sesgo en los estimadores o, equivalentemente, la fuerte dependencia de los estimadores resultantes respecto de las hipótesis consideradas.

En segundo lugar, es imprescindible elegir adecuadamente las covariantes subrogantes. Para ello hará falta tener en cuenta de manera especial, además de criterios estadísticos, criterios inherentes al contenido de las variables (clínicos, epidemiológicos y del mismo proceso de recogida de datos) en la medida en que éstas permitan «aportar» información bien sobre el proceso de no respuesta, bien sobre las covariantes parcialmente observadas. Por estos motivos se hace necesario un análisis de sensibilidad complementario que permita dar interpretaciones razonables de las estimaciones bajo los diferentes supuestos de patrón de no respuesta.

Otra de las limitaciones con las que se encuentra este tipo de análisis se debe al crecimiento geométrico de la dimensión de los parámetros que intervienen en las distintas modelizaciones. Como única alternativa a dicha dificultad se propone una restricción en el número de covariantes de interés y de covariantes subrogantes. También la simplificación de las modelizaciones (*i.e.,* estableciendo relaciones funcionales entre las covariantes) puede reducir la dimensión de los parámetros que intervienen en la optimización. Una vez más, estas relaciones funcionales no podrán ser validadas a partir de los datos observados.

Por último, cabe destacar el elevado coste computacional que supone la ejecución de un análisis completo. El tiempo de estimación de un modelo oscila, según la complejidad de éste, de minutos en los más simples a horas —e incluso días— en los más complejos. Todo ello condiciona fuertemente su utilización y se hace necesario el uso de metodologías menos restrictivas, como por ejemplo la metodología semiparamétrica mencionada en la introdución de este trabajo y en la que los autores están trabajando actualmente.

En resumen, los puntos considerados en esta discusión, y a lo largo del artículo, ponen de manifiesto las dificultades subyacentes en el diseño, implementación e interpretación de la metodología paramétrica en análisis de supervivencia con datos no observados en las covariantes y justifican la investigación y el desarrollo de nuevos procedimientos.

6. AGRADECIMIENTOS

Este trabajo ha sido parcialmente subvencionado por el proyecto DGICYT PB95-0776 del Ministerio de Educación y Ciencia. Los autores agradecen a la Dra. Rotnitzky su colaboración durante los últimos años, al Dr. Caylà y a su equipo del Institut Municipal de la Salut la cesión de los datos que ilustran este trabajo, así como su colaboración, y a los evaluadores anónimos del artículo sus valiosos comentarios y sugerencias en la fase de revisión del mismo.

REFERENCIAS

Baker, S.G. (1994). «Regression Analysis of Grouped Survival Data with Incomplete Covariates: Nonignorable Missing-Data and Censoring Mechanisms». *Biometrics*, 50, 821-826.

Caylà, J.A., Jansà, J.M., Artacoz, L., Plasència, A. and AIDS–TB Group (1993). «Predictors of AIDS in a cohort of HIV–infected patients with pulmonary or pleural tuberculosis». *Tubercle and Lung Disease*, 74, 113–120.

Efron, B. (1994). «Missing Data, Imputation and the Bootstrap». *Journal of the American Statistical Association*, 89, 463–479.

Glynn, R.J., Laird, N.M. and Rubin, D.B. (1993). «Multiple Imputation in Mixture Models for Nonignorable Nonresponse With Follow-ups», *Journal of the American Statistical Association*, 88, 423, 984–993.

Glynn, R.J., Laird, N.M. and Rubin, D.B. (1986). «Selection Modeling Versus Mixture Modeling with Nonignorable Nonresponse», in *Drawing Inferences from Self Selected Samples*, Howard Wainer (editor). Springer-Verlag, New York, 115–157.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis With Missing Data*. John Wiley, New York.

Miller, R.G. (1980). *Simultaneous Statistical Inference*. Springer–Verlag, New York.

Newey, W.K. (1990). «Semiparametric Efficiency Bounds». *Journal of Applied Econometrics*, 5, 99–135.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). «Estimation of Regression Coefficients When Some Regressors Are Not Always Observed». *Journal of the American Statistical Association*, 89, 427, 846–866.

Rotnitzky, A. and Wypij, D. (1994). «A Note on the Bias of Estimators with Missing Data». *Biometrics*, 50, 1163–1170.

Serrat, C. and Gómez, G. (1995). «Estimació de paràmetres amb dades absents en un estudi d'anàlisi de supervivència». *Document de Recerca DR95/10*. Dept. Estadística i Investigació Operativa. Universitat Politècnica Catalunya.

Serrat, C., Gómez, G., García, P. and Caylà, J. (1998). «CD4+ Lymphocytes and Tuberculin Skin Test as Survival Predictors in Pulmonary Tuberculosis HIV-Infected Patients». *International Journal of Epidemiology*, 27, 703-712.

APÉNDICE

a) *Si el patrón de no respuesta es monótono, la condición de MAR*

$$P(R_i = r|L_i) = P(R_i = r|L_{(r)i}), i = 1, \cdots, n \qquad [A]$$

donde $r \in \{0,1\}^K$*, es equivalente a la condición*

$$\begin{aligned} P(R_{ik} = 1|R_{i(k-1)} = 1, L_i) &= P(R_{ik} = 1|R_{i(k-1)} = \\ &= 1, L_{i1}, \cdots, L_{i(k-1)}), i = 1, \cdots, n \end{aligned} \qquad [B]$$

para cada $k = 1, \cdots, K$.

En particular, el proceso de no respuesta es MCAR si, y sólo si, para cada $k = 1, \cdots, K$ *la probabilidad de observación de la* k*-ésima variable condicionada a la observación de la variable* $(k-1)$*-ésima y a los datos potenciales es constante, es decir*

$$P(R_{ik} = 1|R_{i(k-1)} = 1, L_i) = P(R_{ik} = 1|R_{i(k-1)} = 1), i = 1, \cdots, n$$

para cada $k = 1, \cdots, K$.

Demostración: Cuando el patrón de no respuesta es monótono el espacio muestral del vector de respuesta, R_i -de dimensión K-, queda reducido al conjunto

$$\Omega_{R_i} = \{r_0 = (0, \cdots, 0), r_1 = (1, 0, \cdots, 0), r_2 = (1, 1, 0, \cdots, 0), \cdots, r_K = (1, \cdots, 1)\}$$

y la condición [A] se puede reescribir

$$P(R_i = r_k|L_i) = P(R_i = r_k|L_{i1}, \cdots, L_{ik}), i = 1, \cdots, n \qquad [A']$$

para cada $k = 0, \cdots, K$.

– Para probar la implicación [A] $\Rightarrow$ [B], demostraremos por inducción sobre el índice k que, bajo la hipótesis [A′], las probabilidades

$$P(R_{ik} = 1|L_i) \text{ y } P(R_{ik} = 1|R_{i(k-1)} = 1, L_i)$$

sólo dependen de los datos $L_{i1}, \cdots, L_{i(k-1)}$.

En efecto, si $k = 1$, tenemos

$$P(R_{i1} = 1|L_i) = 1 - P(R_{i1} = 0|L_i) = 1 - P(R_i = r_0|L_i)$$

que por la condición [A′] no depende de los datos.

Supongamos demostrado el enunciado hasta un índice j, $(1 < j < K)$. Para demostrar el resultado para el índice $j+1$, es suficiente comprobar las relaciones

(1) $$P(R_{i(j+1)} = 1|L_i) = P(R_{ij} = 1|L_i) - P(R_i = r_j|L_i)$$

(2) $$P(R_{i(j+1)} = 1|R_{ij} = 1, L_i) = 1 - \frac{P(R_i = r_j|L_i)}{P(R_{ij} = 1|L_i)}.$$

Como los segundos términos de las igualdades anteriores dependen como mucho, por la condición [A′] y la hipótesis de inducción, solamente de los datos $L_{i1}, \cdots, L_{ij}$, queda demostrado que la condición [A] implica la condición [B].

La expresión (1) se deduce directamente ya que, por la monotonía del patrón de no respuesta, $\{R_{ij}\}$ es unión disjunta de $\{R_{i(j+1)}\}$ y $\{R_i = r_j\}$.

De manera análoga, la igualdad (2) se obtiene a partir de

$$\begin{aligned} P(R_{i(j+1)} = 1|R_{ij} = 1, L_i) &= P(R_{ij} = 1|R_{ij} = 1, L_i) - P(R_i = r_j|R_{ij} = 1, L_i) = \\ &= 1 - P(R_i = r_j|R_{ij} = 1, L_i) = \\ &= 1 - \frac{P(R_i = r_j, R_{ij} = 1|L_i)}{P(R_{ij} = 1|L_i)} = \\ &= 1 - \frac{P(R_i = r_j|L_i)}{P(R_{ij} = 1|L_i)}. \end{aligned}$$

– Para ver que la condición [B] es suficiente para que el proceso de no respuesta sea MAR basta comprobar que se cumple la condición [A′].

Si $k = 0$, la condición [A′] es cierta ya que $P(R_i = r_0|L_i) = P(R_{i1} = 0|L_i) = 1 - P(R_{i1} = 1|L_i)$, bajo la hipótesis [B], no depende de los datos.

Para $k = 1, \cdots, K$, y utilizando que el patrón de no respuesta es monótono, tenemos

$$\begin{aligned} P(R_i = r_k|L_i) &= P(R_{i1} = 1|L_i) \cdot P(R_{i2} = 1|R_{i1} = 1, L_i) \cdots \\ &\quad \cdots P(R_{ik} = 1|R_{i(k-1)} = 1, L_i) \cdot P(R_{i(k+1)} = 0|R_{ik} = 1, L_i) \end{aligned}$$

y si se cumple [B] todos los factores del segundo término de la igualdad anterior dependen, a lo sumo, de los datos $L_{i1}, \cdots, L_{ik}$, lo que prueba la implicación que queríamos demostrar.

En particular, que las probabilidades $P(R_{ik} = 1|R_{i(k-1)} = 1, L_i)$, $k = 1, \cdots, K$, sean constantes es la condición para que el proceso de no respuesta sea MCAR. Si denotamos por π_k la probabilidad condicionada $P(R_{ik} = 1|R_{i(k-1)} = 1, L_i)$ entonces las probabilidades del espacio muestral Ω_{R_i} son

$$P(R_i = r_k|L_i) = \begin{cases} 1-\pi_1 & \text{si } k=0 \\ \pi_1\pi_2\cdots\pi_k(1-\pi_{k+1}) & \text{si } k=1,\cdots,K-1, \\ \prod_{k=1}^{K}\pi_k & \text{si } k=K \end{cases}$$

que no dependen de los datos L_i.

■

b) *Si el patrón de no respuesta es monótono y los datos son MCAR, los p-valores, $p_1, p_2, \cdots, p_K$, resultantes del estadístico basado en la razón de verosimilitud de H_{0k} contra H_{Ak} en las K pruebas*

$$\begin{aligned} H_{0k} &: logit(P(R_{ik}=1|R_{i(k-1)}=1, L_{i1}, \cdots, L_{i(k-1)})) = \alpha_{k1} \\ H_{Ak} &: logit(P(R_{ik}=1|R_{i(k-1)}=1, L_{i1}, \cdots, L_{i(k-1)})) = \\ &= \alpha_{k1} + \alpha_{k2}^t h_k(L_{i1}, \cdots, L_{i(k-1)}) \qquad k=1,\cdots,K, \end{aligned}$$

siguen una distribución uniforme en $(0,1)$ y son independientes entre sí.

Demostración: Fijado un valor de k, $k=1,\cdots,K$, como el patrón de no respuesta es MCAR la hipótesis H_{0k} es verdadera y $logit(P(R_{ik}=1|R_{i(k-1)}=1)) = \alpha_{k1}$, y por lo tanto

$$P(R_{ik}=1|R_{i(k-1)}=1) = \frac{\exp(\alpha_{k1})}{1+\exp(\alpha_{k1})} = \pi_k.$$

En consecuencia, la variable aleatoria resultante de calcular las frecuencias relativas del suceso «observación de la variable k-ésima condicionada a la observación de la variable $(k-1)$-ésima», $\{R_{ik}=1|R_{i(k-1)}=1\}$, y que denotamos por f_{ik}, se comporta, asintóticamente, como una variable aleatoria $N\left(\pi_k, \sqrt{\frac{\pi_k(1-\pi_k)}{n_k}}\right)$, donde n_k es el número de individuos para los cuales se ha observado la variable $L_{i(k-1)}$ (por construcción, n_1 es el tamaño muestral).

Si p_k es el p-valor resultante del estadístico basado en la razón de verosimilitud de H_{0k} contra H_{Ak}, tenemos $p_k = P(|f_{ik}-\pi_k| > |f_{ik,datos}-\pi_k|)$.

Para demostrar que $p_k \sim U(0,1)$ es suficiente demostrar que $\forall p \in [0,1]$, $P(p_k \leq p) = p$. Y en efecto, fijado p, $0 \leq p \leq 1$, si denotamos por I_{1-p} el correspondiente intervalo con probabilidad $1-p$ centrado en π_k para la distribución $N\left(\pi_k, \sqrt{\frac{\pi_k(1-\pi_k)}{n_k}}\right)$, obtenemos $P(p_k \leq p) = P(f_{ik} \notin I_{1-p}) = p$, como queríamos probar.

Los K p-valores, $p_1, p_2, \cdots, p_K$, son independientes entre sí por el hecho de que las probabilidades condicionadas π_k no dependen de los datos L_i.

■

c) *Si el patrón de no respuesta es MAR, la maximización de la función de verosimilitud no depende de las probabilidades de no respuesta.*

Demostración: De acuerdo con la notación introducida, es suficiente demostrar que la expresión $L(\theta, \psi)$ se puede descomponer en producto de dos funciones $L_1(\theta)$ y $L_2(\psi)$.

En efecto, si las probabilidades de no respuesta $P(R_i = r|L_i; \psi)$ no dependen de los datos no observados $L_{(\bar{r})i}$ entonces

$$\int f_c(L_i;\theta) \cdot P(R_i = r|L_i;\psi) dL_{(\bar{r})i} = P(R_i = r|L_i;\psi) \cdot \int f_c(L_i;\theta) dL_{(\bar{r})i}$$

y la expresión

$$\begin{aligned} L(\theta,\psi) &= \prod_{i=1}^{n} \left\{ [f_c(L_i;\theta) \cdot P(R_i = \mathbf{1}|L_i;\psi)]^{I(R_i=\mathbf{1})} \right. \\ &\quad \left. \prod_{r \neq \mathbf{1}} \left[\int f_c(L_i;\theta) \cdot P(R_i = r|L_i;\psi) dL_{(\bar{r})i} \right]^{I(R_i=r)} \right\} \end{aligned}$$

admite la descomposición $L(\theta, \psi) = L_1(\theta) \cdot L_2(\psi)$ donde

$$\begin{aligned} L_1(\theta) &= \prod_{i=1}^{n} \left\{ f_c(L_i;\theta)^{I(R_i=\mathbf{1})} \cdot \prod_{r \neq \mathbf{1}} \left[\int f_c(L_i;\theta) dL_{(\bar{r})i} \right]^{I(R_i=r)} \right\} \\ L_2(\psi) &= \prod_{i=1}^{n} \left\{ \prod_r P(R_i = r|L_i;\psi)^{I(R_i=r)} \right\}. \end{aligned}$$

■

ENGLISH SUMMARY

SURVIVAL STUDIES WITH NON OBSERVED DATA. DIFFICULTIES CONCERNING THE PARAMETRIC APPROACH

G. GÓMEZ*
C. SERRAT**
Universitat Politècnica de Catalunya

We analyze a parametric approach in survival studies when some of the covariates have not been observed in some individuals. Difficulties in the design, in the computations, as well as philosophicals, that arise in the specification of the likelihood function and in its optimization are discussed. A sensitivity analysis of the estimators, implemented in S-PLUS, is presented. As illustration, the introduced methodology is applied in a survival study in a cohort of pulmonary tuberculosis HIV–infected patients.

Keywords: Incomplete data models, maximum likelihood, missing at random, non-ignorable non-response, survival analysis, validation study.

AMS Classification (MSC 2000): 62F10, 62H99, 92C60.

This work has been partially supported by DGICYT project PB95–0776.

* Departament d'Estadística i Investigació Operativa. Universitat Politècnica de Catalunya. Pau Gargallo, 5. 08028 Barcelona. E-mail: ggg@eio.upc.es.

** Departament de Matemàtica Aplicada I. Universitat Politècnica de Catalunya. Av. Dr. Gregorio Marañón, 44–50. 08028 Barcelona. E-mail: carles@ma1.upc.es.

– Received February 1998.
– Accepted February de 1999.

The missing data problem is already a classical problem that has not been yet solved satisfactorily. Most of the existing methodologies are based on the assumption that non observed data are missing completely at random or at most missing at random. This problem includes those situations where the dependent variable is the survival time which itself could be censored.

A brief review to the different approaches and their advantages and inconvenients follows. First, we could base the analysis on those individuals with all the observed covariates. The inference based on the so-called complete sample is biased and not consistent because the observed individuals are not necessarily a good representation of the overall sample. Furthermore, the analysis based on the complete sample would be inefficient due to reduction in the sample size. Therefore, although this approach is appealing because can be handled with the existing software, its use has to be avoided.

A second approach consists in the imputation of the non observed values. The main problem here relies in that the observed values are used to model the non observed ones and this assumption might not be true. The resulting estimators could then be seriously biased.

A parametric modelization and the solution via maximum likelihood is another possible approach. As it is known this methodology is asymptotically efficient. This is a widely used method; however, its implementation is not straightforward and the corresponding estimators rely heavily on a large number of assumptions that cannot be validated.

The semiparametric approach is a fourth way of handling the missing data problem that allows to model only what is strictly necessary; in our situation we will have to model the relationship between survival and the covariates and the non-response pattern.

In this paper, and previously to an analysis based on a semiparametric approach, we use a completely parametric point of view. This parametric analysis has two main goals: on one hand to show the inconvenients, practical and philosophical, in the specification of the likelihood function and in its optimization, and on the other to design a methodology that would allow to determine how the resulting estimators depend on the non-response pattern.

We start introducing some needed terminology. The potential data vector $L = (L_1, L_2, \cdots, L_K)$ for an arbitrary individual is defined as the vector that contains his/her observed and not observed data. The response vector $R = (R_1, R_2, \cdots, R_K)$ for this individual has k-th ($k = 1, \cdots, K$) component equal to 1 if the k-th variable has been observed and 0 otherwise. L is split into the subvectors $L_{(R)}$ and $L_{(\bar{R})}$ corresponding to the observed and unobserved data, respectively.

Different non-response patterns will depend on the values that $\pi_L(r) = P(R = r|L)$ takes for an arbitrary individual, where $r = (r_1, r_2, \cdots, r_K)$, $r_k \in \{0, 1\}$, $k = 1, \cdots, K$.

The non-response process is Missing Completely at Random (MCAR), if and only if, $\pi_L(r)$ is independent of L. The process is Missing at Random (MAR), if and only if, $\pi_L(r)$ depends at most of $L_{(r)}$. The non-response process is Non-Ignorable (NI) if $\pi_L(r)$ depends on $L_{(\bar{r})}$.

In a survival study the main variable T, is usually the elapsed time between an origin (randomization date in a clinical trial, treatment initiation, etc) and the realization of an event (death, diagnosis of AIDS, etc). Data in these studies is often right censored and the observed data are $Y = min\{T, C\}$ and $\delta = \mathbf{1}\{T \leq C\} = \mathbf{1}\{Y = T\}$ where C is the censoring time. For each individual we also have the values of certain covariates. Let X be the covariates vector. If X_* is a subvector of X and V_* is another set of covariates, we will say that V_* is a surrogate of X_* if X_* and V_* are strongly correlated. The goal in a survival study, with missing data in the covariates, is to model T in terms of the covariates vector X, using the information provided by X and by the surrogate vector V_*.

The sequence of the paper is now the following. We start testing whether the non-response process is MCAR. Then, we introduce a hierarchical scheme to check the sensitivity of the model parameters under different non-response patterns with the objective to elucidate about the non-ignorability of the non-response process. The paper ends with an exhaustive illustration where the methodology is applied and the main disadvantages of the parametric approach are shown.

The test to check whether or not the non-response process is MCAR is based on the comparison of the probabilities $P(R_{ik} = 1), i = 1, \cdots, n$ and $P(R_{ik} = 1|L_i), i = 1, \cdots, n$ for every $k = 1, \cdots, K$. The hypothesis test is formulated as H_0 : MCAR non-response process versus H_A : MAR or NI non-response process, and two different procedures are developed depending on whether or not the non-response process is monotonous.

To develop the parametric approach, let $f_c(l; \theta)$ be the density function for the complete data and $P(R_i = r|L_i; \psi)$ be the probability of observing certain L_i components. We wish to estimate the parameters θ and ψ by maximum likelihood when the likelihood function from the observed data is given by

$$\begin{aligned} L(\theta, \psi) &= \prod_{i=1}^{n} \left\{ [f_c(L_i; \theta) \cdot P(R_i = \mathbf{1}|L_i; \psi)]^{I(R_i=\mathbf{1})} \right. \\ &\quad \left. \prod_{r \neq \mathbf{1}} \left[\int f_c(L_i; \theta) \cdot P(R_i = r|L_i; \psi) dL_{(\bar{r})i} \right]^{I(R_i=r)} \right\}. \end{aligned}$$

It can be proved that if the non-response process is either MCAR or MAR, the maximum likelihood estimator of θ is independent of the non-response process.

In our survival study the density function can be expressed as

$$f_c(L_i; \theta) = f_c((Y_i, \delta_i, V_i^t, X_i^t)^t; \theta) = f_c(X_i; \theta) \cdot f_c(Y_i, \delta_i|X_i; \theta) \cdot f_c(V_i|Y_i, \delta_i, X_i; \theta).$$

To solve the problem parametrically we have to correctly specify: a) the distribution of the covariates, X, b) the distribution of the observed times, Y, conditioned to X, c) the distribution of the surrogate covariates, V, conditioned to Y and to X and d) the non-response probabilities conditioned to the potential data.

To illustrate the above methodology we study survival time from tuberculosis in patients who have contracted the human immunodeficiency virus. The following variables are collected for each individual in the sample: sex, age, risk group, previous tuberculosis treatment, tuberculin skin test (PPD for short), radiological test, bacteriological test, T-CD4+ levels (CD4 for short), ... among others. The main goal in this study is to determine the survival predictive role of CD4 and PPD. The methodological problem relies in the fact that the percentage of missing data in CD4 and PPD are respectively $37{,}5\,\%$ and $50{,}5\,\%$. Following some discussions with the epidemiologists, the following variables are decided to use as surrogates of CD4 and PPD: previous treatment (TR), radiology (RA) and bacteriology (BA). The MCAR validation is carried out and it is concluded that the null hypothesis cannot be rejected. The parametric problem is settled up after modelling: a) the distribution of the main variables, CD4 and PPD, through conditional logistic models, b) the observed survival time, conditioned to CD4 and PPD, via a Weibull model, c) the distributions of the surrogate covariates, TR,RA and BA, conditioned to CD4, PPD and the observed survival time, through the logarithms of the odds ratios, d) the non-response pattern is modelled as conditional logistic probabilities with respect to the covariates in the following way:

$$\begin{aligned}
logit(P(R_{CD4}=1)) &= \alpha_{10}+\alpha_{11}TR+ \\
&+ \alpha_{12}RA1+\alpha_{13}RA2+\alpha_{14}BA1+\alpha_{15}BA2 \\
&+ \alpha_{16}\mathbf{1}\{CD4>14\}+\alpha_{17}PPD, \\
logit(P(R_{PPD}=1|R_{CD4}=1)) &= \alpha_{20}+\alpha_{21}TR+ \\
&+ \alpha_{22}RA1+\alpha_{23}RA2+\alpha_{24}BA1+\alpha_{25}BA2 \\
&+ \alpha_{26}\mathbf{1}\{CD4>14\}+\alpha_{27}PPD \quad \text{y} \\
logit(P(R_{PPD}=1|R_{CD4}=0)) &= \alpha_{30}+\alpha_{31}TR+ \\
&+ \alpha_{32}RA1+\alpha_{33}RA2+\alpha_{34}BA1+\alpha_{35}BA2 \\
&+ \alpha_{36}\mathbf{1}\{CD4>14\}+\alpha_{37}PPD.
\end{aligned}$$

Thus, if θ stands for all the parameters involved in the a) through c) modelization and $\psi=(\alpha_{10},\cdots,\alpha_{37})$ corresponds to those in d), the likelihood $L(\theta,\psi)$ will depend, at most, on a 135 dimension parameter vector.

The way that the different non-response patterns are designed in d) allows us to establish the following five cases:

$$\begin{aligned}
\alpha_{ij} = 0 \quad i = 1,2,3 \quad j = 1,\cdots,7 \quad &\Rightarrow \quad \text{MCAR} \\
\alpha_{ij} = 0 \quad i = 1,2,3 \quad j = 6,7 \quad &\Rightarrow \quad \text{MAR} \\
\alpha_{ij} = 0 \quad i = 1,2,3 \quad j = 7 \quad &\Rightarrow \quad \text{NI1} \\
\alpha_{ij} = 0 \quad i = 1,2,3 \quad j = 6 \quad &\Rightarrow \quad \text{NI2} \\
\text{Without restrictions on the values of } \alpha_{ij} \quad &\Rightarrow \quad \text{NI3}
\end{aligned}$$

Tables 1 and 2 summarize the results. It can be concluded that in general the MAR model is not significatively different from the MCAR model. Therefore, the probability of observing CD4 and PPD does not depend on the surrogate observed values. In all the settings, the NI models are significatively different from the MAR. Thus, the probabilty of observing CD4 and PPD depends on their potential values. Finally, it can be concluded that the positivity in the tuberculin skin test is a better predictor for short-time survival.

The parametric approach developed in this paper is of limited use due to the following considerations. First, the methodology depends on a large number of assumptions that can be quite arbitrary and cannot be validated from the observed data. As a consequence, the estimators might be strongly assumption-dependent. Secondly, the choice of the surrogate covariates is crucial in the sense that they have to be based on clinical and epidemiological considerations. Thirdly, the geometrical growth of the parameter dimension is another limitation of this approach. Finally, the execution of a complete analysis is extremely computationally costly –from minutes to days–.

Summarizing, alternative ways to analyze survival models with missing covariates have to be developed. The authors are presently working on the semiparametric approach.