Ciencia en Desarrollo, Vol. 8 No. 1
ISSN 0121-7488 – Enero-Junio de 2017, pp. 21-28

Estudio del efecto de la imputación de fallas en la estimación de la curva de supervivencia bajo censura a intervalo

Study of the effect of failure imputation in estimating the survival curve under interval censoring

Mario César Jaramillo Elorza [a*]
Carlos Mario Lopera Gómez [b*]

Recepción: 22 de febrero de 2016
Aceptación: 29 de diciembre de 2016

Abstract

Most survival analyzes are based on exact failure times and right censored observations, using methods widely known as the Kaplan-Meier (KM). When the data are interval censored is necessary to use the Turnbull's method to estimate the survival function, but in practice is often used the imputation of failure times in this kind of censorship through the midpoint of the interval, the right end of the interval or generating a random point within the interval using the uniform distribution. This paper studies through simulation the effect of three types of imputation on the estimates of the survival curve compared to the method developed by Turnbull. Different simulation scenarios based on the sample size and the time between visits were analyzed. In all scenarios simulation functions estimated using data imputation differ significantly from the true survival function $S(t)$.

***Key words*:** Survival Analysis, Interval Censoring, Data Imputation.

Resumen

La mayoría de los análisis de supervivencia se basan en tiempos de falla exactos y observaciones censuradas a la derecha, utilizándose métodos ampliamente difundidos como el método de Kaplan-Meier (KM). Cuando los datos presentan censura a intervalo es necesario utilizar el método de Turnbull para estimar la función de supervivencia, sin embargo en la práctica se usa con frecuencia la imputación del tiempo de falla en este tipo de censura a través del punto medio del intervalo (PM), el extremo derecho del intervalo (ED) o generando un punto aleatorio dentro del mismo a través de la distribución uniforme. Este trabajo estudia a través de simulación el efecto de los tres tipos de imputación sobre la estimación de la curva de supervivencia en comparación al método desarrollado por Turnbull. Se analizaron diferentes escenarios de simulación basados en el tamaño de muestra y el tiempo entre visitas. En todos los escenarios de simulación las funciones estimadas usando imputación de datos difieren significativamente de la verdadera función de supervivencia $S(t)$.

***Palabras clave*:** Análisis de Supervivencia, Censura de Intervalo, Imputación de Datos.

[a] Escuela de Estadística, Universidad Nacional de Colombia, Sede Medellín, Colombia.
[*] Autor de correspondencia: mcjarami@unal.edu.co
[b] Escuela de Estadística, Universidad Nacional de Colombia, Sede Medellín, Colombia.
[*] Correo electrónico: cmlopera@unal.edu.co

1. Introducción

El análisis de supervivencia es un conjunto de procedimientos estadísticos para el análisis de datos en los que la variable de resultado es el tiempo hasta que ocurre un evento de interés. La función de supervivencia es quizás la función más importante en los estudios de medicina y salud. Como es usual en el análisis de datos de supervivencia, es de interés estimar la función de supervivencia $S(t)$ y evaluar la importancia de factores potenciales de pronóstico o características individuales, sobre este tiempo de supervivencia.

La gran cantidad de estudios epidemiológicos realizados en enfermedades como el cáncer, entre muchas otras, y la cantidad de estudios longitudinales con desenlaces que involucran el tiempo, demuestran la importancia del análisis de supervivencia. Alternativamente al desenlace de supervivencia o tiempo hasta la muerte, el tiempo puede hacer referencia al momento en que una persona presenta cualquier otro evento. Si el evento se presenta en todos los individuos, se podrían aplicar muchos métodos. Sin embargo, lo habitual es que al final del seguimiento, algunas de las personas no han desarrollado el evento de interés, por lo que el verdadero tiempo trascurrido hasta el evento es desconocido. Además, los datos de supervivencia rara vez se distribuyen de forma "normal", y se componen generalmente de muchos eventos al inicio del seguimiento y los eventos tardíos son relativamente pocos. Estas características de los datos son las que hacen que sea necesario un método especial como el análisis de supervivencia.

Las dificultades específicas relacionadas con el análisis de supervivencia surgen en gran medida por el hecho de que sólo algunas personas han experimentado el evento, por lo tanto, el tiempo de supervivencia se desconoce en un subconjunto de sujetos del estudio. Este fenómeno se llama censura y sus mecanismos pueden ser debido a que el individuo no ha experimentado el desenlace al momento de cierre del estudio; porque se pierde del seguimiento; o porque el sujeto presenta un evento diferente que hace imposible un seguimiento posterior (riesgo competitivo). En este último caso, las censuras deben estimarse de manera diferente y requiere un análisis especial de los datos. Pero al visualizar el proceso de supervivencia de un individuo como una línea de tiempo, pueden verse tres tipos de censuras: Si el evento (suponiendo que llegara a ocurrir) está más allá del final del período de seguimiento, esta situación se conoce como censura a derecha. Otro caso se presenta cuando se observa el evento de interés antes de la primera evaluación, pero no se sabe exactamente cuándo ocurrió. Este tipo de censura es la censura a izquierda. Y por último, el tiempo trascurrido hasta el evento también puede ser censurado en intervalo, cuando los individuos salen y entran del seguimiento (por ejemplo, cuando los individuos se presentan a controles médicos con cierta frecuencia), el individuo presenta el evento de interés al regreso del seguimiento pero la única información que se tiene en este caso, es que el evento se produce dentro de un intervalo de tiempo dado.

La mayoría de los datos de supervivencia incluyen solamente observaciones censuradas a derecha y tiempos de falla exactos, utilizándose métodos ampliamente difundidos como el método de Kaplan-Meier (KM), pruebas de logrank y regresión de Cox (análisis de riesgos proporcionales). Sin embargo, los métodos que soportan datos censurados a izquierda o en intervalo no son tan conocidos. Pocos paquetes estadísticos permiten estos datos, y por esta razón, la práctica común entre los investigadores consiste en simplemente ignorar y descartar las censuras a izquierda de los datos, o realizar una imputación del desenlace para las censuras de intervalo. Es decir, asumir que el evento que ha ocurrido dentro del intervalo $(L_i, U_i]$ ocurrió ya sea en el límite inferior o superior del intervalo o, en el punto medio del mismo. Diferentes autores [1], [2], [3] y [4] manifiestan que asumir el tiempo de supervivencia de intervalo como si fuera exacto puede conducir a estimadores sesgados así como también a conclusiones y estimaciones parciales que no son completamente fidedignas. Estas afirmaciones motivan de alguna manera, a propuestas distintas relacionadas con el tratamiento que se le debe dar a estas censuras, con el fin de evitar sesgos y que se incorpore mayor información. En este trabajo, se pretende estudiar a través de simulación el efecto de varios tipos de imputación sobre la estimación de la curva de supervivencia en

comparación al método de Turnbull para estimación bajo censura arbitraria [5, 6, 7].

En la Sección 2 se presentan los métodos estadísticos utilizados. Un estudio de simulación es presentado en la Sección 3. La Sección 4 recopila los resultados obtenidos. Finalmente, en la Sección 5 se dan algunas conclusiones y recomendaciones con base en los hallazgos encontrados.

2. Métodos

En el desarrollo del trabajo se utilizará el método de estimación de la función de supervivencia, desarrollado por [5], [6] y [7], que incluye los tres tipos de censuras, mediante el algoritmo implementado por [8], para el software R versión 3.1.1 [9].

2.1. Estimador no paramétrico de Turnbull

En los estudios longitudinales, donde los individuos son monitoreados durante un lapso de tiempo prefijado, o visitados periódicamente un cierto número de veces, el tiempo $T_i,\ i = 1,\ldots,n$, hasta que ocurre el evento de interés para cada individuo es desconocido. Sólo se sabe que está dentro de un intervalo entre dos visitas, es decir, entre la visita en el tiempo L_i y la visita en el tiempo U_i con $L_i < T_i \leq U_i$. Si el evento ocurre exactamente en el momento de una visita, lo cual es muy poco probable, pero puede ocurrir, se tiene un tiempo de supervivencia exacto. En este caso se asume que $L_i = T_i = U_i$.

Por otra parte, se sabe que para los individuos cuyos tiempos están censurados a derecha, el evento de interés no ha ocurrido hasta la última visita, pero puede ocurrir en cualquier instante desde ese momento en adelante. Por consiguiente se supone en este caso que T_i puede ocurrir dentro del intervalo $(L_i, +\infty)$, con L_i igual al periodo de tiempo desde el comienzo del estudio hasta la última visita y $U_i = +\infty$.

De modo semejante, para los individuos cuyos tiempos están censurados a izquierda, se sabe que el evento de interés ha ocurrido antes de la primera visita, y por lo tanto, suponemos que T_i ha ocurrido en el intervalo $(0, U_i]$ con $L_i = 0$ representando el comienzo del estudio y U_i es el tiempo hasta la primera visita. El método de Turnbull generaliza cualquier situación con combinaciones de tiempos de supervivencia (exacto o intervalo) y censuras a izquierda y derecha, como datos de supervivencia de intervalo. Por lo tanto, los tiempos de supervivencia exacta, así como también datos de censura a izquierda y derecha, son todos casos especiales de datos de supervivencia con censura de intervalo con $L_i = U_i$ para fallas exactas, $U_i = +\infty$ para las censuras a derecha y $L_i = 0$ para censuras a izquierda.

Como uno de los objetivos principales en análisis de supervivencia, es estimar la función de supervivencia e investigar la importancia de factores potenciales de pronóstico bajo tiempos de supervivencia con censura a intervalo, el número de factores bajo estudio debería depender del propósito del estudio. Como lo sugiere [10], la estimación no paramétrica de la función de distribución acumulada $F(t)$, o en su defecto de la función de supervivencia $S(t)$, es preferible a su estimación paramétrica, por varias razones. Por ejemplo, una elección equivocada de la distribución paramétrica de T podría conducir a conclusiones erróneas de $S(t)$. Además, podría ser difícil encontrar una distribución paramétrica apropiada para ajustar los datos. Hougaard da el ejemplo de tiempos de vida de una población cuya función hazard muestra la llamada forma de bañera: la cual en un principio decrece pocos años, luego permanece constante durante muchos años y por último empieza a aumentar. En este caso, el mejor ajuste probablemente se obtendría de una mezcla de distribuciones.

En el caso de censura a derecha, se podría usar el estimador de Kaplan-Meier para obtener a $S(t)$ [11]. Sin embargo, con datos censurados en intervalo, el método de Kaplan-Meier, no puede ser aplicado, y han sido [5], [6] y [7] quienes han desarrollado el estimador no paramétrico de máxima verosimilitud (NPMLE según siglas en inglés) para estos datos.

El estimador de Turnbull, se basa en una muestra de intervalos observados $[L_i, R_i],\ i = 1,\ldots,n$, los cuales contienen las variables aleatorias independientes $T_1,\ldots,T_n$. Como se mencionó antes, una observación exacta de T_i se da sólo si $L_i = R_i$.

Dado este ejemplo la función de verosimilitud a ser maximizada es la siguiente:

$$L(F) = \prod_{i=1}^{n} [F(R_i+) - F(L_i-)] \qquad (1)$$

Para resolver este problema de maximización [5] define dos conjuntos $\gamma = \{L_i,\ i = 1,\ldots,n\}$ y $\kappa = \{R_i,\ i = 1,\ldots,n\}$ que contienen los extremos izquierdos y derechos de los intervalos, respectivamente. De estos conjuntos se forman nuevos intervalos $[q_1, p_1]$, $[q_2, p_2], \ldots,$ $[q_m, p_m]$, tales que $q_j \in \gamma$, $p_j \in \kappa$ y además no contienen otros miembros de γ y κ, exceptuando a sus puntos extremos.

Se puede probar que una función que maximice (1) es constante entre los intervalos $[q_j, p_j]$ e indefinido dentro de ellos. Note que esto implica que $\widehat{P}(T \in (p_{j-1}, q_j)) = 0$ para cualquier j. Como la función de distribución es no decreciente, la cual no es constante entre los intervalos, puede no maximizar a $L(F)$. Si se denotan los incrementos de la función F dentro de los intervalos $[q_j, p_j]$ como $s_j, j = 1,\ldots,m$, entonces $L(F)$ debe ser maximizada como una función de $s_1, s_2, \ldots, s_m$, sujeto a las restricciones $s_j \geq 0$ y $s_m = 1 - \sum_{j=1}^{m-1} s_j$. Peto aborda este problema de maximización usando el algoritmo de Newton-Raphson. En contraste con Peto, [7], propone el uso del algoritmo de auto-consistencia para el mismo problema de maximización. La idea de éste algoritmo fue presentada primero por [12] y su aplicación para la maximización en (1) es como sigue.

Sea $\alpha_{ij} = I_{\{[q_j,p_j] \in [L_i,R_i]\}}$, $i = 1,\ldots,n$, $j = 1,\ldots,m$, las variables indicadoras que confirman si el intervalo $[q_j, p_j]$ se encuentra dentro o no del intervalo $[L_i, R_i]$, entonces la probabilidad de que T_i se encuentre dentro del intervalo $[q_j, p_j]$ dado un vector $s = (s_1, s_2, \ldots, s_m)'$ está dada por:

$$\mu_{ij}(s) = \frac{\alpha_{ij} s_j}{\sum_{k=1}^{m} \alpha_{ik} s_k} \qquad (2)$$

puesto que $\widehat{F}$ es constante fuera de los intervalos $[q_j, p_j]$. La proporción de observaciones en el intervalo $[q_j, p_j]$ es igual a:

$$\pi_j(s) = \frac{1}{n} \sum_{i=1}^{n} \mu_{ij}(s) \qquad (3)$$

y un vector $s = (s_1, s_2, \ldots, s_m)'$ es llamado auto-consistente, si

$$s_j = \pi_j(s),\ j = 1,\ldots,m.$$

Siguiendo esta definición, el algoritmo de auto-consistencia de Turnbull para el cálculo del estimador no paramétrico de $F(t)$ se puede implementar siguiendo estos pasos:

1. Obtenga estimaciones iniciales de $\mathbf{s}$; por ejemplo, $s_j^{(0)} = \frac{1}{m},\ j = 1,\ldots,m$.
2. Para $i = 1,\ldots,n,\ \ j = 1,\ldots,m$, calcule $\mu_{ij}(\mathbf{s}^{(0)})$ acorde a (2), y luego $\pi_j(\mathbf{s}^{(0)})$ de acuerdo a (3).
3. Obtenga estimaciones mejoradas para $\mathbf{s}$ hallando $s_j^{(1)} = \pi_j(\mathbf{s}^{(0)})$.
4. Retorne al paso 2, reemplazando $\mathbf{s}^{(0)}$ por $\mathbf{s}^{(1)}$ y continúe hasta que se logre la convergencia.

3. Estudio de Simulación

Para establecer el efecto de la imputación de fallas exactas cuando en realidad se tiene una censura a intervalo, sobre la estimación de la función de supervivencia se utilizarán datos de falla lognormales con parámetros fijos para la simulación en valores $\mu = 3{,}78419$ y $\sigma = 0{,}133$, que emulan la historia de eventos en individuos que pueden sufrir fallas entre los 20 y 70 años.

Se asume un punto de partida aleatorio en el que el individuo comienza las visitas en que es evaluado, en donde se registrará si éste tiene o no el evento. Así, se construyen intervalos de tiempo de una de las siguientes formas:

- $(0, U_i]$ un individuo llegó al estudio en el tiempo U_i pero ya tenía el evento de interés (esto constituye una censura a izquierda, la cual se puede ver como una censura a intervalo),
- $(L_i, U_i]$ un individuo llegó al estudio y asistió a visitas regulares, y en el tiempo L_i fue la última visita en la cual no tenía el evento pero

al volver en la siguiente visita (al tiempo $U_i = L_i + \text{TEV}$, con TEV: el tiempo entre visitas) el individuo ya tiene el evento de interés (esto también constituye una censura a intervalo), y

- $(L_i, +\infty)$ un individuo llegó al estudio, asistió a varias visitas regulares, y en el tiempo L_i fue la última visita de la que se tiene registro del individuo en el estudio, sin que éste haya presentado el evento (esto constituye una censura a derecha).

Con este esquema de datos, no se tienen tiempos de falla exactos (aunque también las fallas exactas se pueden considerar como censuras a intervalo con $L_i = U_i$) y todos los datos deben entrarse al análisis como intervalos de tiempo.

Los factores de simulación que se van a variar son:

1. Método de imputación (MI): se estudiarán los casos en que las censuras de intervalo son imputadas a través del punto medio del intervalo (PM), utilizando el extremo derecho del mismo (ED) o generando un valor aleatorio con la distribución uniforme (UN) con parámetros A y B, tomando como valores de éstos, los límites donde se encuentra la falla (tanto para censura a izquierda como para censura de intervalo. Lo cual lleva a tiempos de falla "exactos" y facilita los análisis, ya que la estimación de Kaplan-Meier (KM) para la curva de supervivencia puede ser estimada. Adicionalmente, se considera el caso en que ninguna imputación es llevada a cabo, es decir usando los datos en forma de intervalos de tiempo, lo cual necesariamente lleva a utilizar el estimador de Turnbull (TB) para la función de supervivencia que tiene en cuenta censura arbitraria.
2. Tiempo entre visitas (TEV): indica con que frecuencia los individuos asisten a los controles en el estudio. Interesan valores de TEV $= 1, 2, 4$ y 6 años.
3. Tamaño de la muestra (n): este factor tiene como objetivo establecer el efecto sobre el proceso de estimación del número de individuos en el estudio. Se tomarán valores de $n = 50, 100, 200, 500$.

Para comparar el desempeño de las estimaciones usando el estimador KM, bajo los métodos de imputación $\widehat{S}(t)_{\text{PM}}$, $\widehat{S}(t)_{\text{ED}}$ y $\widehat{S}(t)_{\text{UN}}$, y usando el estimador de Turnbull $\widehat{S}(t)_{\text{TB}}$, se utilizará como control la función de supervivencia real, notada $S(t)$. Esto permite a través de las diferencias observadas entre cada una de las curvas $\widehat{S}(t)_{\text{TB}}$, $\widehat{S}(t)_{\text{PM}}$, $\widehat{S}(t)_{\text{ED}}$ y $\widehat{S}(t)_{\text{UN}}$, y la curva de supervivencia de referencia $S(t)$, establecer el efecto de la imputación sobre la estimación.

Para comparar las curvas de supervivencia resultantes de la simulación, se generan $N = 1000$ muestras independientes para cada uno de los 16 escenarios de simulación (resultantes de las combinaciones de los niveles de los factores TEV y n). Luego en cada escenario, se realizan las estimaciones de la función de supervivencia, de acuerdo al factor de imputación: $\widehat{S}(t)_{\text{TB}}$, $\widehat{S}(t)_{\text{PM}}$, $\widehat{S}(t)_{\text{ED}}$ y $\widehat{S}(t)_{\text{UN}}$, y se comparan con la función de supervivencia de control $S(t)$. Tal comparación se realiza usando el error cuadrático medio integrado (ECMI) como una medida global de error. Para calcular el ECMI con $N = 1000$ simulaciones en cada escenario, se utiliza la siguiente fórmula:

$$\text{ECMI}_i = \frac{1}{N}\sum_{j=1}^{N}\int\left[\widehat{S}_j(t)_i - S(t)\right]^2 dt,$$

donde, $i = \text{TB}, \text{PM}, \text{ED}, \text{UN}$ representa el método de estimación de la función de supervivencia, y $S(t)$ es la función de supervivencia real.

Adicionalmente, para establecer dónde se dan las diferencias entre las curvas de supervivencias estimadas con la real, se calculó el error cuadrático medio (ECM) en la estimación de los cuantiles $q_{0,05}$, $q_{0,1}$, $q_{0,25}$, $q_{0,5}$, $q_{0,75}$, $q_{0,9}$, $q_{0,95}$, de manera que se establece el correspondiente sesgo de estimación de los métodos estudiados (TB, ED, PM y UN). El ECM se calculó para $i = \text{TB}, \text{PM}, \text{ED}, \text{UN}$ y $h = 0{,}05, 0{,}1, 0{,}25, 0{,}5, 0{,}75, 0{,}9, 0{,}95$ como:

$$\text{ECM}_{i,h} = \frac{1}{N}\sum_{j=1}^{N}\left(\widehat{q}_{h,i,j} - q_h\right)^2,$$

donde $\widehat{q}_{h,i,j}$ son $N = 1000$ estimaciones en cada uno de los métodos estudiados $i = \text{TB}, \text{PM}, \text{ED}, \text{UN}$ de los cuantiles reales

q_h, $h = 0{,}05, 0{,}1, 0{,}25, 0{,}5, 0{,}75, 0{,}9, 0{,}95$ de la distribución lognormal con parámetros $\mu = 3{,}78419$ y $\sigma = 0{,}133$.

4. Resultados

4.1. Diferencias en las funciones de supervivencia

La medida de error que se utiliza para comparar las estimaciones basadas en imputación $\widehat{S}(t)_{\text{TB}}$, $\widehat{S}(t)_{\text{PM}}$, $\widehat{S}(t)_{\text{ED}}$ y $\widehat{S}(t)_{\text{UN}}$, con la función de supervivencia verdadera $S(t)$, es el ECMI definido en la Sección anterior. Un valor pequeño del ECMI indica que el método de estimación correspondiente produce una curva de supervivencia estimada que es muy cercana a la curva de supervivencia real a lo largo del tiempo, por el contrario valores altos del ECMI indica que las curvas comparadas tienen diferencias a lo largo del tiempo.

La Tabla 1 muestra los ECMI obtenidos en cada uno de los 16 escenarios de simulación considerados.

Tabla 1. ECMI estimado con los métodos TB, PM, ED y UN

n	TEV	ECMI_{TB}	ECMI_{ED}	ECMI_{PM}	ECMI_{UN}
50	1	3.04	42.11	32.51	23.68
50	2	2.24	30.25	32.04	22.81
50	4	2.50	29.69	31.68	23.25
50	6	2.95	34.16	32.06	23.85
100	1	1.58	41.45	31.85	22.37
100	2	1.18	28.82	31.30	22.24
100	4	1.35	29.28	31.53	21.56
100	6	1.64	33.21	31.24	22.31
200	1	0.87	41.17	30.88	21.51
200	2	0.71	28.82	30.63	21.41
200	4	0.76	28.64	30.86	21.44
200	6	0.96	33.15	30.79	21.75
500	1	0.45	40.73	30.78	21.36
500	2	0.38	28.31	30.52	21.05
500	4	0.42	28.36	30.53	21.31
500	6	0.50	32.74	30.62	21.44

En todos los escenarios de simulación el ECMI muestra que las funciones estimadas $\widehat{S}(t)_{\text{PM}}$, $\widehat{S}(t)_{\text{ED}}$ y $\widehat{S}(t)_{\text{UN}}$ difieren significativamente de $S(t)$, lo cual indica que las estimaciones basadas en estas curvas pueden estar muy alejadas de la realidad. Por otro lado el ECMI asociado a la estimación de Turnbull ($\widehat{S}(t)_{\text{TB}}$) tiene los valores más pequeños en todos los escenarios, lo cual sucede sin importar el tamaño de muestra. Sin embargo a medida que el tamaño de muestra aumenta este error disminuye su valor. En el análisis del tiempo entre visitas (TEV) se puede observar que hay un patrón consistente en todos los valores del tamaño de muestra considerados, que indica que TEV $= 2$ años provoca un ECMI menor que en los demás valores de TEV. Vale la pena resaltar que $\widehat{S}(t)_{\text{UN}}$ resultó mejor que los métodos de imputación $\widehat{S}(t)_{\text{PM}}$ y $\widehat{S}(t)_{\text{ED}}$, lo cual sugiere que un trabajo futuro podría ser cambiar el tipo de distribución o usar métodos bayesianos para la imputación.

La Figura 1 ilustra uno de los escenarios considerados en el estudio de simulación ($n = 500$, TEV $= 2$), donde claramente se observan diferencias entre las curvas de supervivencia estimadas usando los diferentes métodos de imputación y la supervivencia real, mientras que la supervivencia estimada mediante Turnbull se ajusta bien a ésta última.

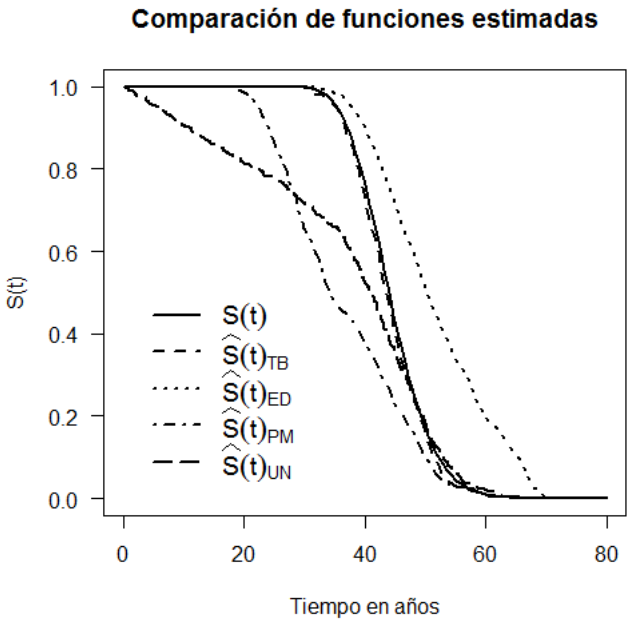

Figura 1. Diferencias entre la curva real y las curvas estimadas mediante Turnbull(TB), KM(ED), KM(PM) y KM(UN). Una realización del caso simulado con $n = 500$ y TEV $= 2$

4.2. Sesgos de estimación de algunos cuantiles

A continuación se presentan los ECM calculados en los métodos estudiados.

Tabla 2. ECM para las estimaciones de los cuantiles $q_h,\ h = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$, usando el método de imputación TB

TEV	$n = 50$						
	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
1	7.04	5.35	4.74	4.19	5.91	9.72	14.48
2	6.63	4.77	3.17	2.85	3.94	6.28	10.42
4	7.42	5.25	3.62	3.33	3.86	5.39	7.21
6	8.89	6.34	4.43	3.86	4.33	6.73	9.13
TEV	$n = 100$						
	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
1	3.98	2.97	2.09	2.24	3.07	5.02	7.71
2	3.51	2.38	1.57	1.51	2.03	3.48	5.33
4	4.56	3.42	1.96	1.68	1.99	2.92	4.40
6	5.30	3.84	2.50	2.08	2.39	3.55	5.47
TEV	$n = 200$						
	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
1	2.26	1.70	1.19	1.27	1.49	2.61	4.18
2	2.15	1.37	0.96	0.92	1.24	1.90	2.92
4	2.57	1.82	1.18	1.07	1.21	1.74	2.61
6	3.17	2.20	1.53	1.41	1.64	2.23	2.94
TEV	$n = 500$						
	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
1	0.96	0.85	0.64	0.67	0.88	1.25	1.88
2	0.99	0.74	0.60	0.58	0.69	0.98	1.41
4	1.25	0.90	0.65	0.63	0.72	0.91	1.22
6	1.55	1.14	0.78	0.79	0.81	1.11	1.54

Tabla 3. ECM para las estimaciones de los cuantiles $q_h,\ h = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$, usando el método de imputación PM

TEV	$n = 50$						
	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
1	165.49	168.62	138.61	85.48	12.65	15.36	17.62
2	164.54	166.63	139.39	83.35	12.42	7.87	12.15
4	167.14	167.28	137.01	85.50	13.10	6.25	6.87
6	167.26	168.03	141.58	87.32	12.80	6.04	7.34
TEV	$n = 100$						
	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
1	165.26	164.17	137.61	88.52	8.12	8.36	13.32
2	164.59	164.02	138.20	84.05	9.93	4.57	6.07
4	164.72	164.11	139.17	80.92	9.89	3.61	4.48
6	163.16	162.85	138.08	83.28	10.19	3.83	4.46
TEV	$n = 200$						
	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
1	166.97	164.47	139.13	90.56	5.64	4.20	8.61
2	167.33	165.09	139.55	85.49	8.35	2.69	3.02
4	166.54	164.57	138.84	86.38	9.23	2.98	2.87
6	166.43	164.34	139.08	86.14	8.74	2.68	2.51
TEV	$n = 500$						
	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
1	169.19	165.61	139.08	94.08	4.56	1.41	3.88
2	168.21	164.81	138.75	87.59	7.81	1.80	1.36
4	168.90	165.40	139.68	88.73	8.46	2.37	1.63
6	168.81	165.33	139.09	88.27	8.04	2.08	1.20

Note que los sesgos de estimación usando el método TB (Tabla 2) son menores que los obtenidos con los métodos de imputación PM, ED y UN (Tablas 3, 4 y 5 respectivamente). En particular, los sesgos de estimación asociados a los métodos de imputación PM y UN (Tablas 3 y 5 respectivamente) son mayores en los cuantiles más pequeños, mientras que para el método de imputación ED (Tabla 4) los sesgos mayores se presentan en los cuantiles más grandes.

Tabla 4. ECM para las estimaciones de los cuantiles $q_h,\ h = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$, usando el método de imputación ED

TEV	$n = 50$						
	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
1	22.52	26.42	44.60	100.98	182.69	200.88	178.32
2	16.88	17.20	28.22	59.37	137.31	182.56	162.60
4	21.98	22.74	33.24	56.38	116.38	172.10	157.15
6	28.24	29.97	42.18	64.89	120.29	174.19	158.02
TEV	$n = 100$						
	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
1	18.70	23.81	42.49	98.48	186.02	207.18	177.33
2	14.85	16.36	25.35	56.98	135.44	183.85	167.64
4	21.79	23.25	31.59	54.11	111.96	170.32	155.11
6	27.38	30.24	40.59	64.35	119.06	173.53	155.97
TEV	$n = 200$						
	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
1	16.76	21.95	41.03	95.70	185.01	209.67	183.92
2	13.87	15.57	25.23	55.81	134.88	185.48	171.75
4	19.80	22.32	31.59	54.00	113.30	171.37	162.32
6	25.44	29.38	39.75	63.97	118.75	173.10	162.46
TEV	$n = 500$						
	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
1	16.04	21.43	40.34	95.71	183.70	210.25	184.77
2	13.29	15.17	24.40	54.38	134.34	186.30	173.79
4	19.32	22.02	31.21	53.50	111.79	169.67	163.49
6	24.71	28.97	39.48	63.51	118.20	172.38	164.02

Tabla 5. ECM para las estimaciones de los cuantiles $q_h,\ h = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$, usando el método de imputación UN

TEV	$n = 50$						
	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
1	852.74	724.81	202.11	11.18	7.52	23.87	34.30
2	841.04	704.97	198.83	13.43	3.79	10.73	21.38
4	853.18	723.58	208.55	14.24	2.95	5.88	11.00
6	850.41	722.33	209.20	16.48	3.36	6.79	12.05
TEV	$n = 100$						
	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
1	848.78	656.58	181.10	7.74	4.43	18.38	29.11
2	843.46	652.62	183.35	8.84	1.92	7.17	15.02
4	832.98	642.38	170.98	7.80	1.41	4.75	6.49
6	844.66	654.18	178.19	9.21	1.52	5.35	6.96
TEV	$n = 200$						
	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
1	886.72	697.47	189.02	6.32	2.53	14.99	26.09
2	883.50	694.42	188.16	7.85	0.89	4.92	11.34
4	882.39	698.06	187.09	7.86	0.76	3.30	5.79
6	888.15	698.12	189.55	8.36	0.77	4.03	6.80
TEV	$n = 500$						
	$q_{0.05}$	$q_{0.1}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$	$q_{0.95}$
1	892.17	689.72	176.75	5.58	1.55	13.11	23.76
2	891.07	687.80	173.09	6.98	0.37	3.90	9.88
4	894.27	688.67	177.89	7.32	0.33	2.11	4.84
6	892.76	689.59	175.63	7.54	0.31	2.93	5.91

Ahora, en general (Tablas 2, 3, 4 y 5) observe que a medida que el tamaño de muestra aumenta los sesgos medidos con el ECM disminuyen, y que los resultados señalan que el tiempo óptimo entre visitas sería de dos años, ya que en este caso los ECM resultaron menores que en los demás valores de este factor.

5. Conclusiones y Recomendaciones

- En los análisis realizados, las curvas de supervivencia estimadas después de realizar una imputación de datos (PM, ED y UN) difieren significativamente de la curva de supervivencia real en todos los tamaños de muestra y en los diferentes TEV, mientras que el método de Turnbull (TB) estima bien en todos los escenarios. También puede concluirse que un tiempo entre visitas igual a 2 años, independiente del tamaño de muestra, es óptimo para estimar la curva de supervivencia, ya que en éste caso se presentaron diferencias más pequeñas que las obtenidas en los escenarios restantes.
- El análisis de los resultados de la estimación de sesgos para los cuantiles $q_h,\ h = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$, usando los métodos TB, PM, ED y UN (Tablas 2, 3, 4 y 5), muestra que en general el método TB presenta menores sesgos en la estimación que los métodos de imputación. También, como es de esperarse a medida que el tamaño de muestra aumenta los sesgos medidos con el ECM disminuyen. Los resultados de las Tablas 3 y 5 establecen que en general los métodos de imputación usando el punto medio del intervalo o un valor aleatorio de la distribución uniforme, afectan las estimaciones de los cuantiles más pequeños, mientras que el método de imputación mediante el extremo derecho del intervalo afecta a los cuantiles más grandes (Tabla 4).
- La imputación de las censuras arbitrarias usando cualquiera de los métodos estudiados resulta en grandes errores, lo cual puede producir resultados sesgados que implicarían por ejemplo en el ámbito clínico un error en el diagnóstico, en el tratamiento y por ende, en el pronóstico de la enfermedad. Por lo tanto se recomienda no realizar imputaciones de los datos censurados a intervalo y en lugar de ello usar el método de estimación de Turnbull que fue diseñado para este tipo de censura.

Referencias

[1] Rucker, G. y Messerer, D. 'Remission duration: an example of interval-censored observation', *Statistics in Medicine* **7**, 1139–1145, 1988.

[2] Odell, P., Anderson, K. y D'agostinho, R. 'Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model', *Biometrics* **48**, 951–959, 1992.

[3] Dorey, F. J., Little, R. y Schenker, N. 'Multiple imputation for threshold-crossing data with interval censoring', *Statistics in Medicine* **12**, 1589–1603, 1993.

[4] Iceland J. *The Dynamics of Poverty Spells and Issues of Left-Censoring. No. 97-378*, PSC Research Report Series January 1997, 1997.

[5] Peto, R. 'Experimental survival curves for interval-censored data', *Journal of the Royal Statistical Society, Series C* **22**, 86–91, 1973.

[6] Turnbull, B. W. 'Nonparametric estimation of a survivorship function with doubly censored data', *Journal of the American statistical association* **69**(345), 169–173, 1974.

[7] Turnbull, B. W. 'The empirical distribution function with arbitrarily grouped censored and truncated data', *Journal of the Royal Statistical Society, Series B* **38**(3), 290–295, 1976.

[8] Giolo, S. 'Turnbull's Nonparametric estimator for interval-censored data', *Department of Statistics, Federal University of Paraná* pp. 1–10, 2004.

[9] R Development Core Team. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014.

[10] Hougaard, P. 'Fundamentals of survival data', *Biometrics* **55**, 13–22, 1999.

[11] Kaplan, E. L. y Meier, P. 'Nonparametric estimation from incomplete observations', *Journal of the American statistical association* **53**(282), 457–481, 1958.

[12] Efron, B. 'The two sample problem with censored data', *University of California Press* pp. 831–853, 1967.