METODOS DE SUPERVIVENCIA PARA ESTUDIAR LA DESERCIÓN

Rubén Belmonte Coloma

Introducción

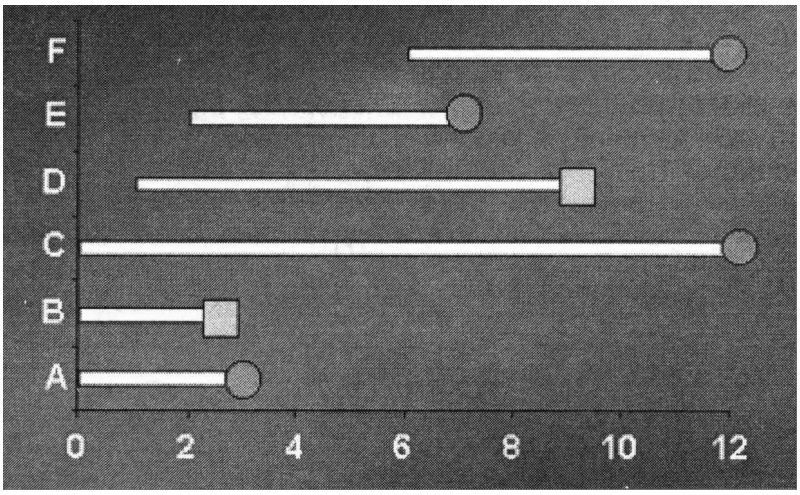
El análisis de supervivencia es una técnica estadística que tiene innumerables aplicaciones en el área de la salud, en el área actuarial, en la planificación industrial, para citar algunas. En este artículo se pretende establecer algunas bases para invitar a estudiar un fenómeno académico cuyo origen puede ser de preocupación y este es el de la deserción académica en la universidad.

Para nadie es desconocido el hecho de que la deserción de universitarios en el sistema público es un fenómeno de múltiples causas, económicas, sociales, principalmente, sin embargo este problema no es objeto de este artículo, se dan las bases del tratamiento técnico del tiempo de permanencia.

Se denomina análisis de supervivencia al conjunto de técnicas que permiten estudiar la variable "tiempo hasta que ocurre un evento" y su dependencia de otras posibles variables explicatorias. En este ejemplo, en el estudio de la deserción, el tiempo hasta que ocurre la deserción (tiempo de supervivencia) sin estudiar, por ahora, su dependencia a los factores externos. Debido a que la variable tiempo es una variable continua podría ser, en principio, estudiada mediante las técnicas de análisis de la varianza o los modelos de regresión. Hay, sin embargo, dos dificultades importantes para este planteamiento. En primer lugar, en la mayor parte de los estudios la variable tiempo no tiene una distribución normal, más bien suele tener una distribución asimétrica y aunque podrían intentarse transformaciones que la normalizaran, existe una segunda dificultad que justifica un planteamiento específico para estas variables, y es que para observarlas se tiene que prolongar el estudio durante un período de tiempo suficientemente largo, en el cual suelen ocurrir pérdidas, que imposibilitan la observación del evento.

Existen tres motivos por los que pueden aparecer estas pérdidas, en primer lugar por fin del estudio. Supóngase, por ejemplo, que para evaluar se sigue en el tiempo, durante un año, a dos grupos de alumnos. A los de un grupo se les aplica algún método que incentive la permanencia en la universidad y a los de otro no, y se registró la duración del intervalo de tiempo entre la intervención pedagógica (o la entrada en el estudio, para el grupo no intervenido) y la deserción. Al final del estudio puede haber individuos que no hayan desertaron. Algunos de los individuos, y puede ser un número importante, cambian de carrera o simplemente están matriculados pero no asisten desaparecerán del estudio en algún momento del mismo por diversos motivos. En estudios de supervivencia en el área de la salud Aunque los ejemplos anteriores son del ámbito académico los más significativos están en el área de salud, estos mismos problemas aparecen en cualquier estudio que necesite un largo tiempo de observación.

Hay que tener en cuenta también que la variable es el tiempo hasta que ocurre un evento, y está definida por la duración del intervalo temporal entre los instantes en que empieza la observación y ocurre el evento. En los ejemplos citados, la observación no comienza en el mismo instante para todos los individuos. En algunos textos se denomina pérdida por la izquierda a esta no coincidencia de los tiempos en que comienza la observación, ya que, si el estudio está diseñado para acabar en un tiempo determinado, el efecto de esta no coincidencia es reducir, para los que empiezan más tarde, el tiempo de observación. En el esquema de la figura se detallan todas las posibles pérdidas. Evidentemente, se pueden evitar las pérdidas por la izquierda diseñando el estudio para que acabe, no en un tiempo establecido con carácter general, sino, para cada individuo, en un tiempo determinado después del inicio de la observación.



Esquema temporal de un estudio para observar tiempos de espera para un evento, por ejemplo supervivencia en una intervención quirúrgica. Con el círculo se representan las pérdidas y con el cuadrado las deserciones (ocurrencia del evento). El individuo A desaparece del estudio 3 años después del ingreso (sería una pérdida en sentido estricto). El B deserta a los 2,5 años (cinco semestres. El C sigue en la universidad hasta el final del estudio (sería una pérdida a los 12 semestres por fin del estudio). El D, al que ingresa después de un semestre, deserta en el 9, el tiempo de supervivencia sería 8 semestres (hay 1 semestre de pérdida por la izquierda). El E, al que se le ingresa en el segundo semestre, se pierde en el 7 (sería una pérdida a los 5 meses, ya que hay pérdida en sentido estricto y pérdida por la izquierda). El F, al que ingresa en el sexto semestre, sigue en la universidad al acabar el estudio, sería una pérdida a los 6 meses (existe pérdida por fin del estudio y pérdida por la izquierda).
Si se quisiera aplicar un modelo de regresión lineal a un estudio de este tipo, habría que eliminar del mismo las observaciones perdidas, ya que para ellas no se conoce el valor de la variable; sin embargo sí se tiene alguna información útil sobre la misma: se sabe que es mayor que el tiempo en el que se produjo la pérdida.

Distribución de la variable tiempo de espera

La variable tiempo de espera es una variable aleatoria continua y no negativa, cuya función de probabilidad puede especificarse de varias maneras. La primera es la habitual función densidad de probabilidad f(t), y relacionadas con ella, la función de supervivencia S(t) y la función de riesgo h(t).

La función densidad de probabilidad f(t) para una variable continua se define como una función que permite calcular la probabilidad de que la variable tome valores en un intervalo a través de la fórmula:

$$P(a < T < b) = \int_a^b f(t)dt \qquad 0 < t < \infty$$

La función de supervivencia S(t) se define como:

$$S(t) = P(T \geq t) = \int_t^{h(t)} f(U)dU$$

Por lo tanto, la función de supervivencia da la probabilidad complementaria de la habitual función de distribución acumulativa F(t) = P(T £ t), es decir S(t) = 1 - F(t).

Otro modo de expresar la probabilidad para la variable tiempo de espera es por medio de la función de riesgo h(t) que es la función de densidad de probabilidad de T, condicionada a que T ^{3} t. Por ejemplo, para la supervivencia al ingreso, la función de riesgo a los 2 años es la de densidad de probabilidad de desertar a los 2 años del ingreso. Esta probabilidad sería, realmente, la que en cada momento le importa al universitario.
Se puede demostrar que

$$h(t) = \frac{f(t)}{S(t)}$$

A veces se usa también la función de riesgo acumulada H(t), más difícil de interpretar, que se define como

$$H(t) = \int_{\square}^{t} h(x)dx$$

y que verifica

$$H(t) = -\ln\big(S(t)\big)$$

Es decir, las cuatro funciones están relacionadas; si se conoce una cualquiera de ellas, se pueden obtener las demás.
A pesar de que el tiempo es una variable continua, un observador sólo tiene acceso a valores discretos de la misma. Los datos observados para cualquiera de las experiencias descritas en la introducción son una serie de valores discretos. Conviene, por lo tanto, definir las funciones anteriores en el caso (práctico) de considerar a la variable tiempo como discreta, es decir, como un conjunto discreto de valores $t_1 < t_2 < ...$ El suponerlos ordenados de menor a mayor no representa ninguna pérdida de generalidad, de hecho es así como se observa el tiempo.

Para una variable discreta, la función densidad de probabilidad f (t) se define como:

$$f(t_i) = P(T = t_i), i = 1, 2, ...$$

y la función de supervivencia:

$$S(t_l) = \sum_{l/t_l \geq t_i} f(t_i)$$

La función de supervivencia da, por lo tanto, para cada valor t_i de T, la probabilidad de que la variable T sea mayor o igual que t_i (en este caso no es la complementaria de la función de distribución puesto que la probabilidad de que T sea igual a t_i, que en las variables discretas en general no es cero, está incluida en ambas funciones), aunque otros textos, justamente para que siga siendo la complementaria de la función de distribución la definen sin incluir el igual.

Las funciones de riesgo y riesgo acumulado para una variable discreta también son:

$$h(t_l) = \frac{f(t_l)}{S(t_l)} \qquad H(t_i) = -\operatorname{Ln} S(t_l)$$

El método Kaplan Meier

El método de Kaplan Meier es un método no paramétrico (no asume ninguna función de probabilidad) y por máxima verosimilitud, es decir se basa en maximizar la función de verosimilitud de la muestra. Una muestra aleatoria de tamaño n, extraída de una población, estará formada por k (k £ n) tiempos $t_1 < t_2 < ... < t_k$ en los que se observan eventos. En cada tiempo t_i existen n_i "individuos en riesgo" (elementos de la muestra para los que el evento puede

ocurrir, o que T ^{3} t_i) y se observan d_i eventos. Además en el intervalo $[t_i,t_{i+1})$ se producen m_i pérdidas.

Se puede demostrar que la función de verosimilitud para toda la muestra es:

$$L = \prod_{i=1}^{k} {h_i}^{d_i} (1 - h_i)^{n_i - d_i}$$

Para construir esta función se ha asumido que la información contenida en las pérdidas es que, para cada una de ellas, el evento ocurre en un tiempo mayor que el tiempo en que se observa la pérdida. Maximizando esta función se encuentra que el estimador de la función de riesgo es

$$\hat{h}_i = \frac{d_i}{n_i} \; i = 1, 2, \ldots, k$$

y para la función de supervivencia, el denominado estimador producto límite o de Kaplan-Meier:

$$\hat{S}(t_i) = \prod_{i|t_i<t_i} \left(1 - \frac{d_i}{n_i}\right)$$

Se sigue en el tiempo a 12 individuos que ingresaron a la universidad y se encuentran los siguientes tiempos de supervivencia en años: 6*, 6, 6, 6, 10, 12*, 12, 15, 15*, 17, 22, 22, donde el asterisco indica deserción; es decir se perdieron 3 individuos en los tiempos 6, 12 y 15. La manera más cómoda de calcular los estimadores anteriores es disponer los datos en una tabla como la que sigue:

tiempo	ind. en riesgo	eventos	F. riesgo	F. supervivencia
6	12	3	3/12=0,25	1
10	8	1	1/8=0,125	0,750
12	7	1	1/7=0,143	0,656
15	5	1	1/5=0,2	0,562
17	3	1	1/3=0,333	0,450
22	2	2	2/2=1	0,300

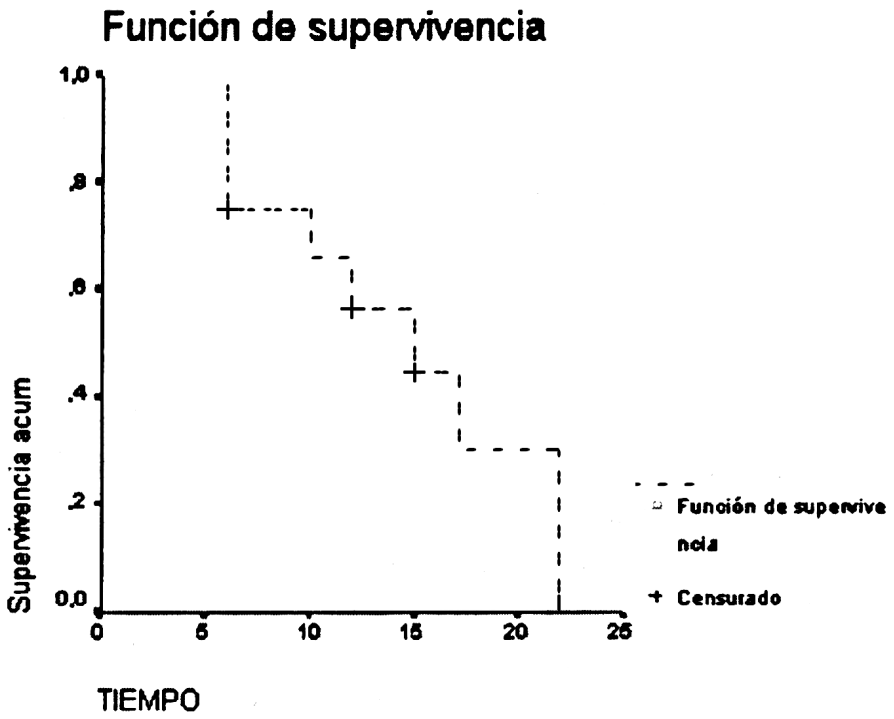
Para analizar estos datos con un paquete estadístico, por ejemplo el SPSS, hay que introducir dos variables: el tiempo y el “status” con un código que indique si en ese tiempo se ha producido el evento o es una pérdida. La “salida” es:

Análisis de supervivencia por Tiempo

Tiempo	Estado	Supervivencia acumulada	Error Estándar	Sucesos acumulados	Renumeración
6	1			1	11
6	1			2	10
6	1	,7500	,1250	3	9
6	0			3	8
10	1	,6563	,1402	4	7
12	1	,5625	,1482	5	6
12	0			5	5
15	1	,4500	,1555	6	4
15	0			6	3
17	1	,3000	,1605	7	2
22	1			8	1
22	1	,0000	,0000	9	0

Numero de casos 12 Censurados: 3 (25,00%) Sucesos : 9

En la tercera columna ("Supervivencia acumulada") aparece la función de supervivencia (S(t) en todos los tiempos en los que ocurren eventos. Esta función se suele representar en una gráfica como



El SPSS también calcula y representa la gráfica de la función de riesgo acumulada (que en su versión en español denomina "Impacto").

E.T.Lee, Statistical Methods for Survival Data Analysis Lifetime, Learning Publications. 1980.
D.R. Cox D Oakes, Analysis of Survival Data, Chapman and Hall. 1984

Ida Tarbell, The Ways of Woman
Regocíjate, pues debajo de nubes y estrellas
nuestro planeta es más que maine o Texas.
Bendice los grandiosos hechos de tener
doce meses, nueve musas y dos sexos,
y en los señoríos de la tierra, infinidad de artes, climas, maravillas de ideas.
Phyllis McGinley, “In Praise of Diversity”

No obstante, la estadística no debe hacerse para demostrar una idea preconcebida
Florence Nihtingale, nota en Physique Sociales, de A. Quetelet

Ahora se había dado cuenta que no sabía nada fundamental y, como un monje agobiado con la conciencia del pecado, se lamentó: “¡Si, ¡y si tan sólo pudiera recordar la estadística!”

Casi todos los estados han devuelto sus censos. Le envío los resultados, con tinta negra, si están basados (en la medida de lo posible) en los datos reales, y con tinta roja si no son los datos regresados, aunque bastante conocidos. Con un pequeño margen para omisiones, son más de cuatro millones, aunque de hecho sabemos que las omisiones han sido muy grandes.
Thomas Jefferson, carta de David Humpreys

Personalmente, nunca me han interesado la ficción ni los cuentos. Lo que me gusta leer son hechos y estadísticas de cualquier tipo; aunque sean hechos acerca de cultivo de rábanos, me interesan. Precisamente ahora, por ejemplo, antes de que usted entrara señala una enciclopedia que se encuentra en un librero estaba leyendo un artículo sobre matemáticas perfectamente puras.

“Mi conocimiento matemático termina en “12 por 12”, pero disfruté inmensamente con ese artículo. No entendí una palabra, pero los hechos, o lo que el hombre cree que son los hechos, parecen siempre encantadores. Este matemático creía en sus hechos, también yo. Primero obtenga sus propios hechos y aquí la voz disminuye hasta casi imperceptible luego puede distorsionarlos tanto como desee”.

Mark Twain, citado por Rudyard Kipling, en from Sea to Sea
No hay remedio más efectivo para aplacar la rabiosa susceptibilidad de la gente que las cifras. Para estar segura, debe prepararse cuidadosamente, abarcar el caso y no sólo su cuarta parte, y debe reunirse por sí misma, no para comprobar alguna teoría. Este tipo de preparación puede encontrarla en el caso nacional.

Índice