Intervalos de confianza anormalmente amplios en regresión logística: interpretación de resultados de programas estadísticos

Jokin de Irala,[1] *Rafael Fernandez-Crehuet Navajas*[2] *y Amparo Serrano del Castillo*[3]

RESUMEN *Este estudio describe el comportamiento de ocho programas estadísticos (BMDP, EGRET, JMP, SAS, SPSS, STATA, STATISTIX y SYSTAT), al realizar una regresión logística con una base de datos simulados en la cual existe un problema numérico creado por la presencia de una celda con frecuencia igual a 0. Los programas responden de manera heterogénea a este problema. La mayor parte de ellos ofrecen señales de alarma, aunque muchos presentan, simultáneamente, resultados incorrectos entre los cuales destacan los intervalos de confianza que tienden al infinito. Estos resultados pueden desorientar al usuario. Se describen diferentes criterios orientativos para detectar estos problemas en situaciones de análisis reales y se recuerda la importancia de la interpretación crítica de los resultados de programas estadísticos.*

La regresión logística es un método de análisis estadístico comúnmente utilizado en epidemiología. Es una técnica atractiva porque permite, con relativa sencillez, estimar valores de lo que en castellano se ha llamado razón de oportunidades, razón de momios, razón de posibilidades, razón de probabilidades o razón de productos cruzados (*odds ratio*, en inglés) (*1*). La razón de posibilidades (RP) es una medida de asociación para respuestas categóricas. La RP es importante en epidemiología porque representa una estimación del riesgo relativo cuando este último no puede estimarse directamente (*2*–*5*).

Para cada sujeto de un estudio, la regresión logística estima la probabilidad Π de obtener la respuesta de interés, condicionada a los valores de las variables independientes para dichos sujetos y siguiendo un modelo específico (*6*). La fórmula empleada para este propósito es la siguiente:

$$\prod = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n}}$$

Donde β_0 es la constante del modelo, β_n, los coeficientes de las variables independientes del modelo y x_n, las variables independientes. En la actualidad, los paquetes estadísticos estiman los coeficientes beta de este modelo y sus correspondientes errores estándar, mediante iteraciones que resuelven ecuaciones de verosimilitud. Una vez obtenidos estos valores, es posible estimar diferentes RP y sus intervalos de confianza (IC). Para una variable cualitativa, los programas pueden calcular la RP de la presencia de una respuesta en una categoría de dicha variable respecto a la categoría de referencia, a condición de que esta variable cualitativa se codifique utilizando el método marginal (*6*, *7*). Con

[1] Las solicitudes de separatas deben enviarse a este autor a la siguiente dirección postal: Universidad de Córdoba, Facultad de Medicina, Departamento de Medicina Preventiva y Salud Pública, Córdoba 14004, España, o al número de fax (349) 57-218278.
[2] Hospital Reina Sofía, Departamento de Medicina Preventiva, Córdoba, España.
[3] Universidad de Córdoba, Facultad de Medicina, Departamento de Medicina Preventiva y Salud Pública, Córdoba, España.

las variables continuas, es posible estimar la RP de la presencia de la respuesta por cada incremento de interés de la variable continua (por ejemplo, por cada 10 unidades de incremento de la variable continua) (*6*, *7*).

El grado de sofisticación del programa estadístico utilizado determinaría la complejidad de los resultados que le ofrece al usuario, pero los programas generalmente presentan, como mínimo, los coeficientes beta y sus errores estándar. Con estos resultados el usuario puede estimar la RP, calculando la exponencial de cada coeficiente y obtener sus IC correspondientes a partir de los errores estándar, cuando el programa no le ofrece dichos datos automáticamente (*6*, *7*).

A pesar de la facilidad con la que estos programas proporcionan resultados, es imprescindible comprender adecuadamente los métodos subyacentes a la regresión logística. Un área de relativa importancia que, frecuentemente, causa perplejidad al usuario de la regresión logística es el de los problemas numéricos. Entre ellos, cabe destacar la inestabilidad del modelo como consecuencia del tamaño muestral inadecuado; el problema de la "separación completa", cuando todos los sujetos con valor de la variable respuesta igual a 1 se pueden separar perfectamente de aquellos sujetos con valor igual a 0, utilizando sus características; el problema de la colinearidad, y, por último, el problema de perfiles con frecuencia igual a 0. Este trabajo se centra en este último caso y, más concretamente, en el problema de la presencia de una celda con frecuencia igual a 0 en un contexto de variables independientes categóricas. Este problema se puede intuir pensando en el ejemplo sencillo del cálculo de la RP a partir de una tabla tetracórica de un hipotético estudio de casos y controles. Los valores de las cuatro celdas serían: *a* (presencia del factor de riesgo y presencia de enfermedad), *b* (presencia del factor de riesgo y ausencia de enfermedad), *c* (ausencia del factor de riesgo y presencia de enfermedad) y *d* (ausencia tanto del factor de riesgo como de la enfermedad). En el caso de que las celdas *b* o *c* tuvieran una frecuencia igual a 0, la RP no podría estimarse cuando se utiliza el método de la razón de productos cruzados (*ad*/*bc*), porque obtendríamos un valor infinito. Lo mismo puede ocurrir en las regresiones logísticas que se realizan sobre bases de datos reales. Los programas estadísticos difieren en su modo de tratar este problema. Es frecuente que los programas presenten al usuario resultados incorrectos. Por tanto, es menester llamar la atención sobre este hecho para que dichos resultados no se publiquen.

Los objetivos de este trabajo son evaluar y describir el comportamiento de ocho programas estadísticos, cuando existen celdas con frecuencia igual a 0 en una base de datos; presentar los signos de alarma que pueden avisar al usuario de la existencia de problemas numéricos y, con esta descripción, recordar la importancia de la interpretación crítica de los resultados que se obtienen con los programas estadísticos.

MATERIALES Y MÉTODOS

El presente estudio se realizó con una base de datos simulados, que se resumen en el cuadro 1. Dicha base consta de 60 sujetos, una variable dependiente y dicotómica, codificada 0 y 1, para la ausencia y presencia de respuesta, respectivamente, y una variable independiente cualitativa x con tres categorías. En el modelo de regresión logística, esta variable cualitativa queda representada por dos variables indicadoras, x_2 y x_3, que se crearon siguiendo el método marginal y representando a las categorías 2 y 3, respectivamente. La categoría de referencia es la primera. El grupo 3 no presenta sujetos con la respuesta 0 (celda con frecuencia igual a 0). Esta base de datos se ha utilizado en ocho programas estadísticos: BMDP, EGRET, JMP, SAS, SPSS, STATA, STATISTIX y SYSTAT. Con cada programa se realizó una regresión logística de la variable y sobre la variable cualitativa x. Con los programas estadísticos se construyó un modelo de regresión y se intentó estimar los coeficientes de las variables del modelo y sus errores estándar. Dicho de otro modo, se trató de alcanzar la "convergencia", es decir, la resolución de ecuaciones de verosimilitud mediante iteraciones. A continuación, se decribe cómo los diferentes programas resuelven las situaciones comparables a la de esta base de datos simulada.

CUADRO 1. Base de datos simulados. La celda (3, 0) tiene una frecuencia igual a cero: no existe ningún sujeto de la categoría "3" que posea el atributo "0". Número total de sujetos = 60

		Variable *Y*		
		0	1	
Variable *X*	1	13	7	20
	2	8	12	20
	3	0	20	20
		21	39	60

RESULTADOS

La presencia de una celda con frecuencia igual a 0 equivale a realizar una operación en la cual el resultado no es definible, no es posible estimar coeficientes o errores estándar, ni tiene sentido estimar una RP para la categoría de la variable cualitativa con frecuencia igual a 0. Debido al problema numérico que se ha introducido, los programas fallan en el intento, no alcanzan la convergencia ni obtienen resultados correctos. Por el contrario, al no poder alcanzar la convergencia, interrumpen su proceso (a veces se observa el mensaje *estimation terminated*). A partir de este momento, los programas difieren entre sí, y presentan resultados que se resumen en el cuadro 2. Todos los programas evaluados excepto STATISTIX produjeron algún tipo de mensaje de advertencia, generalmente indicando que no se pudo alcanzar la convergencia, y en el caso del programa SPSS, que el proceso se había interrumpido por un descenso demasiado pequeño del logaritmo de verosimilitud. Sin embargo, ninguno especificó que la situación se debiese a una celda con frecuencia igual a 0. El programa SAS se caracteriza por indicar con más detalle y mediante el símbolo # que la estimación del coeficiente de la variable x_3 es en realidad una estimación "infinita", es decir, indeterminable (38.21#). STATA avisa que ha abandonado la variable indicadora x_3.

Los programas BMDP, SPSS y STATISTIX ofrecen los resultados que habitualmente calculan (coeficientes β, error estándar e intervalos de confianza de la RP), mientras que JMP y SYSTAT omiten el intervalo de confianza de la RP. Por el contrario, EGRET, SAS y STATA no presentan los errores estándar de los coeficientes ni los intervalos de confianza de la RP. STATA tampoco presenta una estimación para el coeficiente de la categoría problemática de la variable cualitativa *x*.

Algunos programas estiman la RP para la constante y su correspondiente

CUADRO 2. Resultados obtenidos, con los diferentes programas estadísticos, al aplicar la regresión logística a la base de datos simulados

Programas[a] Variables del modelo	Coeficientes	Error estándar	RP	Intervalos de confianza (95%)	Mensaje de advertencia
BMDP LR (1990, PC)					
constante	–0,62	0,469	0,54	(0,20, 1,38)	Sí
x_2	1,03	0,654	2,79	(0,77, 10,04)	"No convergencia"
x_3	10,82	22,300	0,50 E+05	(0,2E–14, 0,1E+25)	
EGRET (1992, 0.03, PC)					
constante	–0,62	?	0,54	No	Sí
x_2	1,03	?	2,79	No	"No convergencia"
x_3	33,89	?	0,52 E+15	No	
JMP (3.01, MAC)					
constante	–0,62	0,469	0,54	(.,.)	Sí
x_2	1,03	0,654	2,79	(0,79, 10,5)	Convergencia por objetivo,
x_3	12,80	99,840	3,70 E+05	(.,.)	no por gradiente
SAS (1987, 6.04, PC)					
constante	–0,62	0,469	No	No	Sí
x_2	1,03	0,654	No	No	"No convergencia"
x_3	38,21#	"."	No	No	
SPSS (1990, 4.01, PC)					
constante	–0,62	0,469	No	No	Sí
x_2	1,03	0,654	2,79	No	"Estimación interrumpida"
x_3	10,82	36,730	0,50 E+05	No	
STATA (1990, 3.10, MAC)					
constante	—	—	No	No	Sí
x_2	1,03	0,654	2,79	(0,77, 10,04)	No se utiliza la variable x_3
x_3	—	—	—	—	No hay resultados
STATISTIX (4.0, PC)					
constante	–0,62	0,469	No	No	No
x_2	1,03	0,654	2,79	(0,77, 10,04)	
x_3	10,19	16,210	2,65 E+04	(0,00, 0,2E+18)	
SYSTAT (1991, 2.00, MAC)					
constante	–0,62	0,469	No	No	Sí
x_2	1,03	0,654	2,00	(0,77, 10,04)	"No convergencia"
x_3	16,82	447,458	0,20 E+08	(0,.)	

[a] Programa (año, versión, PC/Macintosh).
No = no calculados habitualmente por el programa.
Los signos "#", "?" y "." son los presentados por los programas estadísticos correspondientes.
"#" = estimación infinita.
"—" = en blanco.
RP = razón de posibilidades o, en inglés, *odds ratio*.

intervalo de confianza, si bien estas estimaciones carecen de sentido.

DISCUSIÓN

Es importante resaltar que los programas estadísticos que analizan datos con problemas numéricos como los citados en este trabajo en realidad no alcanzan la convergencia. Dichos resultados dependen únicamente del momento en que el programa interrumpe su proceso matemático iterativo, es decir, del criterio de convergencia preestablecido. Esto explica que los programas hayan podido encontrar diferentes valores del coeficiente de x_3 (véase el cuadro 2). Cuantos más intentos realice el programa, mayor será el valor del coeficiente de x_3 y las subsiguientes estimaciones, con una tendencia hacia el infinito.

Como las señales de advertencia observadas en estos análisis no son específicas, es posible que su existencia e importancia pasen desapercibidas al usuario poco avezado en el empleo de estos programas. Para este último, los programas que ofrecen resultados, aunque sean incorrectos (BMDP y STATISTIX), pueden ser los más desorientadores. Los programas que no presentan los intervalos de confianza de la RP de la variable problemática, pero sí los valores de su coeficiente y error estándar (JMP, SPSS y SYSTAT), son menos desorientadores. Sin embargo, aún existe el inconveniente de que los investigadores puedan calcular la RP, calculando la exponencial de dicho coeficiente. Con el error estándar, también podrían calcular los IC. Ambos procedimientos carecerían de sentido en este contexto. Los programas como EGRET, SAS y STATA serían los más apropiados para el usuario sin experiencia, ya que no es posible estimar parámetros para la variable con problemas numéricos.

Las diferencias puestas de manifiesto entre estos programas estadísticos no plantearían problema alguno al usuario que trabaja habitualmente con la regresión logística. Sabría captar cualquier señal de alarma presente en los resultados. Este observaría las advertencias escritas o "notas" que acompañan a las respuestas. La existencia de un coeficiente anormalmente alto (del orden de las decenas), acompañado de un error estándar también de esa misma magnitud, puede ser un signo útil para sospechar un problema numérico. Empero, estos signos no son específicos del problema de celdas con efectivos nulos, sino que también se encuentran en otras situaciones como el de la presencia de colinearidad. La ausencia del cálculo de parámetros como la RP y su IC en un programa que suele realizar estas estimaciones también constituye una señal de alerta. Cuando las estimaciones del programa tienden al infinito, es preciso rechazarlas. Además, hay que hacer hincapié en que estos resultados no deben publicarse. El problema debe solucionarse antes de utilizar ese modelo.

También es preciso recordar que pueden surgir problemas numéricos cuando dos variables tienen algunas categorías con frecuencias pequeñas y dichas variables se introducen conjuntamente en un modelo, como es el caso de un término de interacción. Al construir la variable de interacción, producto de ambas, pueden aparecer celdas con frecuencia nula y los problemas numéricos subsiguientes. Para evitar este problema, se aconseja que en el estudio siempre haya suficiente número de sujetos con cada atributo de las variables cualitativas empleadas (*6, 7*).

Algunos autores han insistido en la importancia que reviste realizar un análisis univariante de los datos antes de efectuar el multivariante, con objeto de poder conocer mejor la calidad y naturaleza de las variables utilizadas y tener una idea general de las asociaciones existentes a nivel univariante (*8*). Asimismo, realizando un análisis descriptivo y univariante con todas las variables antes de realizar el multivariante se podrían detectar variables o categorías que quedan invalidadas por presentar frecuencias insuficientes. En ese caso se debe proceder a eliminar dicha variable o a la agrupación de varias de sus categorías, siguiendo un sentido biológico, para obtener menos grupos y frecuencias más altas en cada uno de ellos. En cualquier caso, es recomendable no quedarse satisfechos con modelos obtenidos automáticamente con paquetes estadísticos. Por el contrario, es preciso evaluar siempre su bondad de ajuste antes de aceptar su validez (*9*). Es preferible utilizar los programas que ofrecen señales de alarma específicas o que no impriman los valores de los estimadores de los coeficientes cuando se den circunstancias como las de los problemas numéricos mencionados en este artículo.

REFERENCIAS

1. Bautista LE. "Razón relativa" y "tasa relativa" como traducciones de *odds ratio* y *hazard ratio* [carta]. *Bol Oficina Sanit Panam* 1995;119:278-280. Tapia JA. Respuesta. *Bol Oficina Sanit Panam* 1995;119:280–282.
2. Rothman KJ. *Modern epidemiology*. 6a ed. Boston: Little, Brown and Co; 1986.
3. Hanley J. Utilizaciones adecuadas del análisis multivariante. *Rev Salud Publica* (Barcelona) 1989;1:45–74.
4. Hennekens CH, Buring JE. *Epidemiology in medicine*. 6a ed. Boston: Little, Brown and Co; 1987.
5. Kleinbaum DG, Kupper LL, Muller KE. *Applied regression analysis and other multivariate methods.* 2a ed. Boston: PWS-KENT; 1988.
6. Hosmer DW. *Computer analysis of health sciences data. PH744 course*. Amherst, MA: University of Massachusetts; 1992.
7. Hosmer DW, Lemeshow SA. *Applied logistic regression.* New York: John Wiley; 1989.
8. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. *Ann Int Med* 1993;118: 201–210.
9. Hosmer DW, Lemeshow SA, Taber S. The importance of assessing the fit of logistic regression models. *Am J Public Health* 1991;81:1630–1635.

Manuscrito recibido el 18 de diciembre de 1995 y aceptado para publicación en versión revisada el 22 de abril de 1996.

ABSTRACT

Abnormally broad confidence intervals in logistic regression: interpretation of results from statistical programs

This study describes the behavior of eight statistical programs (BMDP, EGRET, JMP, SAS, SPSS, STATA, STATISTIX, and SYSTAT) when performing a logistic regression with a simulated data set that contains a numerical problem created by the presence of a cell value equal to zero. The programs respond in different ways to this problem. Most of them give a warning, although many simultaneously present incorrect results, among which are confidence intervals that tend toward infinity. Such results can mislead the user. Various guidelines are offered for detecting these problems in actual analyses, and users are reminded of the importance of critical interpretation of the results of statistical programs.